Predicting the Risk of Lung Cancer in Never-smokers

Cameron Thompson, BSc (honours)

In partial fulfillment of the requirements for the degree of

Master of Science in Applied Health Sciences

(Health Sciences)

Faculty of Applied Health Sciences

Brock University, St. Catharines, Ontario

ABSTRACT

Despite being considered a disease of smokers, approximately 10-15% of lung cancer cases occur in never-smokers. Lung cancer risk prediction models have demonstrated excellent ability to discriminate cases from non-cases, and have been shown to be more efficient at selecting individuals for future screening than current criteria. Existing models have primarily been developed in populations of smokers, thus there was a need to develop an accurate model in never-smokers. This study focused on developing and validating a model using never-smokers from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Cox regression analysis, with six-year follow-up, was used for model building. Predictors included: age, body mass index, education level, personal history of cancer, family history of lung cancer, previous chest X-ray, and secondhand smoke exposure. This model achieved fair discrimination (optimism corrected c-statistic = 0.6645) and good calibration. This represents an improvement on existing never-smoker models, but is not suitable for individual-level risk prediction.

KEY WORDS: Lung cancer, risk prediction, epidemiology, never-smokers, public health

ACKNOWLEDGEMENTS

I'd like to begin by thanking my supervisor, Dr. Martin Tammemagi, for the continuous support of my work over the past two years. Thank you for giving me the chance to pursue research in graduate studies, for providing an amazing opportunity to take on a great project, and for instilling a sense of rigor and professionalism in my work. I'll leave grad school a much better researcher and epidemiologist than I was when I entered, and I'd like to think that a lot of that is due to your advice and feedback.

I'd also like to thank my committee members, Dr. Brent Faught & Dr. Jian Liu. No complete thesis document is truly a solo-project, and so I am very appreciative of your feedback throughout the process. Furthermore, I'd like to thank Dr. Madelyn Law, not just for being the chair at my defence, but for giving me my start in research three years ago. I can safely say that I wouldn't be where I am today if not for my experiences in the I-EQUIP program.

I'd like to thank Brock University for providing a second home for the past eight years, and to Brock and the Ontario Graduate Scholarship for the financial support, allowing me to go the last two years without a "real job". I'd also like to thank Starbucks for the endless nights of free wifi and workspace. Camping out with a large coffee was often the source of some of my best work.

To my parents and the rest of my family: I have no idea why you thought going back for a second degree seemed like a good idea…or a third degree. Thank you for encouraging me to pursue whatever goals I had, for providing endless support over the past 25 years, for proof-reading when I needed it, for allowing me to stay in school

without pressure to jump into the real world, and for accepting that “I dunno” is usually the only answer to “What did you do today?” as a grad student. Thanks.

Finally, to my friends – both at Brock and elsewhere. You guys are the reason that grad school has been the most amazing experience. Thanks for providing a place to bounce ideas off of, to vent about anything, or just for being awesome people to hang out with. Thanks for heated ping pong battles, trail runs around campus, and possibly too frequent Merch nights. It’s been fun.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AJCC – American Joint Committee on Cancer

AUC – Area under the Curve

BEIR – Biologic Effects of Ionizing Radiation

BMI – Body Mass Index

CARET – Beta-Carotene and Retinol Efficiency Trial

CMS – Centres for Medicare & Medicaid Services

COPD – Chronic Obstructive Pulmonary Disorder

CT – Computed Tomography

CXR- Chest X-ray/Chest Radiography

DQX – Dietary questionnaire

EGFR – Epidermal Growth Factor Receptor

EGFR-TK – Epidermal Growth Factor-Tyrosine Kinase

ERT – Estrogen Replacement Therapy

GWAS – Genome Wide Association Studies

HPV – Human Papillomavirus

HR – Hazard Ratio

HRT – Hormone Replacement Therapy

LCINS – Lung Cancer in Never Smokers

LDCT – Low-dose Computed Tomography

LRR – Lifetime Relative Risk

MFP – Multivariable Fractional Polynomial

MICE – Multiple Imputation by Chained Equations

NCI – National Cancer Institute

NCI-SEER – National Cancer Institute Surveillance, Epidemiology and End Results

NDI – National Death Index

NIH – National Institutes of Health

NLST – National Lung Screening Trial

NNS – Number Needed to Screen

NPV – Negative Predictive Value

NSCLC – Non-small Cell Lung Cancer

OR – Odds Ratio

PAH – Polycyclic Aromatic Hydrocarbons

PLCO – Prostate, Lung, Colorectal, and Ovarian

PPV – Positive Predictive Value

RCS – Restricted Cubic Splines

ROC-AUC – Receiving Operating Characteristic – Area under the Curve

RR – Relative Risk

SEER- Surveillance, Epidemiology and End Results

SCLC – Small Cell Lung Cancer

SEP – Socioeconomic Position

SES – Socioeconomic Status

SHS – Secondhand Smoke

SQX – Supplemental Questionnaire

TB - Tuberculosis

TERT – Telomerase reverse transcriptase

TK-1 – Tyrosine Kinase 1

TP53- Tumour Protein 53

UICC – Union for International Cancer Control

USDHHS – United States Department of Health and Human Services

WHO – World Health Organization

CHAPTER I: INTRODUCTION

One of the best-understood and described exposure-disease relationships is that of tobacco smoking and lung cancer. It is estimated that 85-90% of all lung cancer cases are attributable to tobacco smoke (Samet et al., 2009), although this statistic varies geographically and by gender. Relatively less attention has been paid to the other 10-15% of cases, those that occur in individuals who identify themselves as a "never-smoker". With regards to lung cancer research, a never-smoker is an individual who reports smoking less than 100 cigarettes in their lifetime (Yang, 2011). While 10-15% of cases may not seem like an overwhelmingly large number, given the overall incidence of lung cancer in the population it still presents an important public health concern (Thun et al., 2008).

Both epidemiologically and physiologically speaking, lung cancer in never-smokers (LCINS) is considered a distinct disease from its counterpart in smokers, with a different set of risk factors (Sun, Schiller, & Gazdar, 2007). A number of different risk factors have been proposed, however the effect size and overall significance of each individual risk factor typically varies from study to study, and among different populations. Among the most consistently established risk factors are: secondhand smoke (SHS), occupational exposures (asbestos, silica), environmental exposures (radon), gender, family history of lung cancer, previous history of non-malignant lung disease and race/ethnicity (McCarthy, Meza, Jeon, & Moolgavkar, 2012; Rudin et al., 2009; Subramanian & Govindan, 2007).

The most common histological presentation of LCINS is as an adenocarcinoma, with estimates of between 47 and 76% of all cases (Subramanian, Velcheti, Gao, &

Govindan, 2007). This is a significant increase from the overall estimated 40% of all lung cancers (smokers and never-smokers) that are classified as adenocarcinoma (American Cancer Society, 2014). Additionally, never-smokers with lung cancer typically exhibit different molecular mutations than smokers' lung cancers, with a higher frequency of Epidermal Growth Factor Receptor (EGFR) mutations. This has been linked to improved treatment through specific forms of therapy, specifically the use of EGFR inhibitors as part of a chemotherapy regimen (Rudin et al., 2009). While never-smokers or former-smokers are typically diagnosed at an older age than current-smokers (68.7 years vs. 65.4 years)(Tammemagi, Neslund-Dudas, Simoff, & Kvale, 2004), they demonstrate improved 5-year survival (23% vs. 16%) (Nordquist, Simon, Cantor, Alberts, & Bepler, 2004).

Statistical predictive modeling has become common practice for lung cancer, as it has for a number of other cancers and illnesses. However to date there remains a lack of accurate predictive model specifically targeted at never-smokers. Beginning with Bach *et al.*(2003), many predictive models have demonstrated good ability for discerning individuals with cancer from those without, however the majority of these models have either focused solely on smokers, or included them along with former and never-smokers. Due to the overwhelming nature of smoking as a risk factor for lung cancer (OR >20 when comparing current to never-smokers) (Clément-Duchêne et al., 2010), models including current or former smokers are often at risk of washing-out any effects of other predictors. Thus, the aim of this thesis was to develop and validate a risk prediction model for lung cancer in never-smokers, focusing on risk factors not related to active tobacco smoking, with the goal of correctly identifying individuals at the highest risk of developing the disease.

1.1 Lung Cancer in Never-Smokers as a Public Health Concern

Lung cancer is the second most common neoplasm in both men and women in Canada and the United States (Howlader et al., 2011; Statistics Canada, 2015), with an estimate of 26,600 new Canadian cases in 2015 and 224,100 new American cases in 2014 (American Cancer Society, 2014; Statistics Canada, 2015). Even conservatively estimating LCINS as 10% of all cases, this would mean approximately 2,600 new cases of lung cancer occurring in Canadian never-smokers and 22,000 new cases of lung cancer occurring in American never-smokers. Between 2007 and 2011, the National Cancer Institute's Surveillance, Epidemiology and End Results (NCI-SEER) program reported the overall incidence of lung cancer as 60.0 new cases/100,000 person years (72.1 in males, 51.1 in females) (Howlader et al., 2011). This is slightly higher than the incidence of 51.9 new cases/100,000 person years (57.6 in males, 47.5 in females) reported by the Canadian Cancer Society in 2015 (Statistics Canada, 2015). Incidence rates for LCINS vary between cohorts and studies, however they are often similar to those of myeloma in men or cervical cancer in women (Wakelee et al., 2007). When considered a separate disease, LCINS would rank as the 7th leading cause of cancer-related death worldwide (Samet et al., 2009).

1.2 Identifying and Screening High-Risk Individuals

Prognostic outcomes associated with lung cancer can be greatly improved by early detection. Of particular interest is the use of low-dose computed tomography (CT) scanning, which has demonstrated efficacy. When compared to radiography, CT scans have been shown to reduce lung cancer-specific mortality by 20% when included as part of a systematic screening program (Aberle et al., 2011). When detected early, lung cancer

has a reported 5-year survival rate of between 60 and 70%, however only approximately 30% are detected at an early stage where surgery is still possible (Hocking et al., 2010). This demonstrates the importance of an effective screening strategy for lung cancer, as outcomes are drastically improved with earlier detection.

Screening the entire population is not a feasible option, both from a cost standpoint and a risk-benefit standpoint (Aberle et al., 2011). Thus, prediction models should be used to identify the subpopulation at the highest risk, those that are most likely to benefit from screening such as CT scans. A recent model, demonstrating good discrimination and calibration, was developed by Tammemagi and colleagues in 2013. This model estimates the six-year absolute risk for a given individual for developing lung cancer based on a number of risk factors; notably age, race, education level, body mass index (BMI), previous history of cancer, family history of lung cancer, diagnosis of chronic obstructive pulmonary disorder (COPD), and a number of predictors related to smoking status and intensity (Tammemagi et al., 2013). The authors claim that by screening the individuals with a six-year lung cancer diagnosis probability above approximately 0.0095, as identified by their model, they could correctly identify 90% of lung cancer cases for screening. To reduce cost or intervention-associated risk further, they estimate that 80% of cases could be identified by screening individuals above a 0.0151 risk probability – the 65^{th} percentile of smokers risk (Tammemagi et al., 2013).

1.3 Risk Prediction Models & Gaps in Knowledge

To date, a number of risk prediction models have been developed with the goal of identifying high-risk individuals. Beginning with a model developed by Bach and colleagues (2003), risk prediction models have consistently demonstrated an ability to

predict 1-, 5- and occasionally 10-year risk of developing lung cancer. Many of these models looked at predictors such as age, gender, race, socioeconomic status (SES) and body mass index (BMI), as well as a number of different environmental and occupational exposures (Spitz et al., 2007; Tammemagi et al., 2011). However, due largely to the well-known association between tobacco and lung cancer, the vast majority of predictive models contain a number of smoking-related predictors, such as duration and intensity. Given the overwhelming association of smoking and lung cancer (10-20x increased risk), other predictors such as radon or asbestos exposure may appear insignificant in a population including exclusively smokers, or a combination of smokers and non-smokers (Clément-Duchêne et al., 2010; Tammemagi et al., 2011).

Current risk prediction models focusing on never-smokers may be less than ideal for a number of reasons, particularly a lack of consistent information regarding risk factors aside from age and gender (McCarthy et al., 2012). Studies designed to look at associations between single risk factors – such as specific occupational hazards or exposures – and LCINS often have limited sample size and thus low statistical power (Hu, Mao, Dryer, & White, 2002; Lagarde et al., 2001). One study, by Spitz and colleagues (2007) stratified by smoking status and included never-smokers, however this model demonstrated only modest discrimination (Harrell's c-statistic = 0.59). Furthermore, an ever-smoker model adapted and applied to never-smokers did not achieve the same level of discrimination as its ever-smoker counterpart (area under the curve (AUC) = 0.66 in never-smokers) (Tammemagi et al., 2014). Thus, there is a need to develop a risk prediction model in never-smokers that includes a wide range of variables

that improve prediction and demonstrates a higher degree of predictive performance – as measured by discrimination and calibration (to be discussed in a forthcoming section).

1.4 Response to Gaps in Knowledge

In order to address the gaps in knowledge, research was conducted using data from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The PLCO screening trial is a large, randomized prospective screening trial, conducted by the National Cancer Institute (NCI), which recruited individuals aged 55-74 years. Included within this trial's dataset are variables pertaining to cause and time of death (if deceased), diagnosis, staging and histopathology of different cancers, demographic information, anthropometric measures at different time points of life, family history of lung cancer, comorbidities and select medical histories. For this study, data were limited only to those individuals who identified themselves as "never-smokers". Model building was accomplished using Cox survival analysis, with "lung cancer diagnosis" as the outcome variable. Predictors were included based on *a priori* knowledge of LCINS risk factors along with some exploratory investigation of novel variables – those that have not been included in previous models, such as ibuprofen usage - and hazard ratios (HRs) were used to determine the risk associated with each individual predictor. Internal validation was achieved using bootstrapping resampling methods. Descriptive statistics were used to describe the distribution of predictors overall and by lung cancer status.

1.5 Study Aim

As was previously mentioned, the aim of this study was to develop and validate an accurate lung cancer risk prediction model in never-smokers. This was accomplished using high-quality data and sophisticated statistical modeling techniques.

The methodological approach consisted of using Cox survival analysis in Stata 13 statistical software (StataCorp. 2013. *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP). Predictor variables were included via intelligent selection based on *a priori* knowledge of potential risk factors. Non-linear effects of continuous variables were modelled using multivariable fractional polynomials (MFPs) to better fit the true nature of the relationship. Discrimination (ability to identify an individual with cancer from an individual without) was evaluated using Harrell's c-statistic, a measure of how well Cox models correctly identify individuals that have the outcome of interest. Calibration (agreement between expected and observed probabilities) was evaluated using Brier score, calibration plots and mean versus 90^{th} percentile absolute error. Internal validation was conducted using bootstrapping methods to estimate the amount of over-fit to noise in the model. This study aimed to improve on previous models (Tammemagi et al., 2014) in a number of different ways. First, the PLCO dataset contains a larger number of LCINS cases than most other studies (n=276), as well as a large total number of never-smokers (n=69,272), which provided more statistical power than other studies. Also, through the use of MFPs, we aimed to more accurately model the relationship between individual predictors and the risk of developing lung cancer.

1.6 Conclusion

This thesis served to address an important gap in the literature and a public health need regarding LCINS. This predictive model intended to accurately discern never-smoking individuals at the highest risk of developing lung cancer, and thus could provide justification for screening which may lead to early diagnosis and timely treatment. The efficacy of the Tammemagi (2013) model as a tool to identify high-risk groups has lent credence to the importance of developing a similar model specifically for never-smokers. Taking into account that LCINS is the 7th leading cause of cancer-related mortality worldwide (Samet et al., 2009), it is of importance that a predictive model is developed for this disease similar to those for breast cancer or overall lung cancer. For this reason, the development and validation of an accurate risk prediction model is an important step towards addressing LCINS as a public health concern.

CHAPTER II: LITERATURE REVIEW

In order to fully understand the importance and goals of this research, it is important to have a thorough grasp of work that has previously been done in this area. To begin, the biologic and histologic nature of lung cancer must be understood, including the pathologic mechanisms and different subtypes. Next, the burden of the disease on the population, both in general and specific to never-smokers, is detailed. A comprehensive examination of known and suspected risk factors for lung cancer in never-smokers (LCINS), including secondhand smoke (SHS), asbestos exposure, and family history of cancer, sheds light on the nature of potential predictors used to address the study aim. This is followed by an overview of research done in the field of lung cancer screening, particularly low-dose computed tomography (CT) scans, as they pertain to early detection and improved outcomes. Finally, the chapter concludes with a review of different predictive models which have proven effective. This provides evidence of the efficacy of lung cancer risk predictive modelling as a tool, as well as rationale behind the forthcoming Methodology section. Overall, this chapter should demonstrate the importance of the research aim, in addition to providing a basis for future chapters.

2.1 Biology of Lung Cancer

2.1.1 Etiology, Histology & Pathogenesis

Although often talked about as one disease, lung cancer is much more complex than that. Described by Sharma *et al.* (2007) as "a conglomeration of diseases of diverse etiology", there is no singular pathway or mechanism associated with all forms of lung cancer. Carcinogenesis involves irregular cell growth, without proper cell death (apoptosis), with the possibility of invading other tissue (American Cancer Society,

2014). The typical pathway involves the activation of oncogenes or inactivation of tumour suppressor genes through the action of some exogenous or endogenous carcinogenic substance (chemicals in secondhand smoke, environmental toxins, human papilloma virus (HPV), etc.). Among the most commonly mutated genes associated with different types of lung cancer are Epidermal Growth Factor Receptor (*EFGR*) or *KRAS*, which act to control cell proliferation (Damjanov, 2012; Rudin et al., 2009) or tumour protein 53 (*TP53*), a tumour suppressor protein which helps to regulate gene expression (Shigematsu et al., 2005; Wakelee et al., 2007). This results in the uncontrolled and unregulated growth of cells resulting in the formation of tumours (Damjanov, 2012). Different mutations are more commonly associated with different subtypes of lung cancer (Carney, 1995), however the exact cause of many of these is not fully known (Damjanov, 2012).

Lung cancer is classified into two broad types based on the histological nature of the tumour; small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) (Hoffman, Mauer, & Vokes, 2000). NSCLC is further divided into the following sub-types: adenocarcinoma, squamous cell carcinoma and large-cell carcinoma (Hoffman et al., 2000).

2.1.1.1 Small-cell Lung Cancer

Defined by the World Health Organization (WHO) as "a malignant epithelial tumour consisting of small cells with scant cytoplasm" (World Health Organization, 2004), SCLC account for approximately 13% of all lung cancers. The vast majority (~90%) of these cases occur in smokers (van Meerbeeck, Fennell, & De Ruysscher, 2011). Surgical resection is typically not part of the normal course of treatment for

patients with SCLC, who instead undergo rounds of chemotherapy in order to slow the progression of the disease (Damjanov, 2012). Individuals with SCLC usually experience a quicker disease progression and poorer survival outcomes (Damjanov, 2012; van Meerbeeck et al., 2011). However, given the relatively low number of never-smokers who experience SCLC, most LCINS research tends to focus primarily on NSCLC (Subramanian & Govindan, 2007).

2.1.1.2 Non-small-cell Lung Cancer

2.1.1.2.1 Large-cell Carcinoma

Tumours classified as large-cell carcinoma are often centrally located within the lung, and are distinguished by poor differentiation. They are typically diagnosed based on the lack of any identifiers associated with SCLC, squamous-cell or adenocarcinoma (Damjanov, 2012; World Health Organization, 2004). Large-cell carcinoma are the least common of the three major NSCLC sub-types, accounting for approximately 10% of all lung cancer cases (Damjanov, 2012). In past studies, it has been shown to be less common among never-smokers than current or former smokers (Muscat & Wynder, 1995), accounting for approximately 6.2% and 7.7% of all lung cancers in never-smoking males and females respectively. Large-cell carcinoma, like other NSCLCs, is typically treated with a combination of radiation and surgical resection at early stages, and radiation and chemotherapy at more advanced stages (Damjanov, 2012).

2.1.1.2.2 Squamous-cell Carcinoma

Another centrally-located cancer is the squamous-cell carcinoma, which "start in early versions of squamous cells…flat cells that line the inside of the airways in lungs"

(American Cancer Society, 2014). Patients with squamous-cell carcinoma experience distant metastases less commonly than those with other types of lung cancer (World Health Organization, 2004). They account for 25-30% of all lung cancers, however they are much more common in smokers than non-smokers, as approximately 90% of cases are linked to tobacco smoking (American Cancer Society, 2014; World Health Organization, 2004). While squamous-cell carcinoma makes up a relatively large proportion of all lung cancers, a 2006 study by Toh *et al.* demonstrated that only 5.9% of all LCINS cases within their study population were of this variety (Toh et al., 2006). The typical course of treatment is similar to that of large-cell carcinoma, which is standard for all NSCLCs (Damjanov, 2012).

2.1.1.2.3 Adenocarcinoma

Adenocarcinoma of the lung is a subtype consisting of tumours originating from secretory cells located in the epithelial lining of the lung. They typically grow slower than other types of NSCLCs, however metastases are more common than in squamous-cell carcinomas (American Cancer Society, 2014; World Health Organization, 2004). They are the most common form of lung cancer, accounting for approximately 40% of all cases (American Cancer Society, 2014; Damjanov, 2012). Despite already being the most common type overall, adenocarcinoma has been demonstrated to represent an even greater proportion of LCINS. Toh *et al.* (2006) found that adenocarcinoma represented 69.9% of LCINS compared to just 39.9% of lung cancer in current smokers. Adenocarcinoma has a higher prevalence amongst women, individuals of East Asian descent, and younger populations (American Cancer Society, 2014). The increasing incidence amongst never-smokers has led to a number of studies examining the

mechanistic differences between carcinogenesis of adenocarcinoma in never-smokers, former-smokers and current smokers (Nordquist et al., 2004). Treatment for adenocarcinoma is standard for all NSCLCs (Damjanov, 2012), however recent studies have demonstrated positive response rates for *EGFR* tyrosine kinase (*EGFR-TK*) inhibitors (Rudin et al., 2009), as EGFR mutations have shown to be positively associated with adenocarcinomas in never-smokers (Rudin et al., 2009; Sharma et al., 2007).

2.1.2 Clinical Features

2.1.2.1 Symptoms

Damjanov (2012) broadly classifies lung cancer symptoms as being related to one of the following four things: "bronchial irritation or obstruction, local extension of the tumour into the mediastinum or pleural cavity, distant metastases, or systemic effects of the neoplasia." According to the Canadian Cancer Society, symptoms that should lead to consulting a physician include a chronic, worsening cough, chest pain, wheezing, loss of appetite, frequent chest infections or blood in the sputum, amongst others (Canadian Cancer Society, 2014).

2.1.2.2 Diagnosis

According to the American Cancer Society, "most lung cancers do not cause any symptoms until they have spread too far to be cured…", indicating that by the time medical attention is sought, the cancer is already at a more advanced stage (American Cancer Society, 2014). Once identified, the diagnosis of cancer is made using bronchoscopy and associated brush or sputum cytology for centrally located cancers, and biopsy for peripheral cancers (World Health Organization, 2004). It is estimated that

only approximately 10-15% of lung cancer patients are asymptomatic at diagnosis, having them detected during a routine chest X-ray (CXR) or similar examination (Damjanov, 2012). The majority of patients present with an advanced stage of the disease, at which point the prognosis is poor, and treatment often shifts towards management rather than curing (Midthun, 2013). Early stage NSCLC is associated with greatly improved 5-year survival rates following surgical resection (70% for Stage I vs. 50% for Stage II), which highlights the importance of earlier diagnosis (Midthun, 2013). Lung cancer screening programs have been implemented with the goal of earlier detection and reduced mortality. The techniques and outcomes of lung cancer screening programs will be discussed in this thesis.

2.1.3 Tumour Staging

Lung tumours are staged according to a number of anatomical features that are determined following the diagnosis of cancer. This staging is used to determine the extent of the cancer, while providing the physician with some guidance regarding treatment plans and the patients' prognosis (National Cancer Institute, 2013). The most widely used system of tumour staging is the TNM system, which is currently on its 7th edition, and is developed by the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC) (National Cancer Institute, 2013). The TNM system uses three anatomic factors to assign the tumour to a stage group; the T – the size and local extension of the primary tumour, the N – the degree of metastases to regional lymph nodes, and the M – the presence of distant metastases (American Joint Committee on Cancer, 2010).

These T, N and M classifications are used to group tumours into prognostic stage groups (Stage 0 – IV), with increasing group number corresponding to more extensive spread of the disease. Stage 0 is reserved for *in situ* carcinoma only, while Stages I-III correspond to varying degrees of large tumours, or those that have spread to surrounding lymph nodes or organs/tissues. Finally, Stage IV consists of cancers that have spread to distant organs or tissues, and thus have very poor prognosis (National Cancer Institute, 2013).

2.2 Descriptive Epidemiology

The effect of a disease on a population is accurately portrayed through effective presentation of epidemiological data – often a combination of multiple measures (Adami, Hunter, & Trichopoulos, 2008). Of these measures, the most pertinent to predictive modelling is incidence – new cases of a disease occurring over a specified time period. This allows for meaningful comparisons between different groups and populations, while focusing solely on the development of the disease, not the prognosis.

2.2.1 Lung Cancer Statistics

We use statistics such as incidence and mortality rates in lung cancer research to determine the nature of the problem. That is, how pressing is the issue of lung cancer amongst the grand scheme of public health. Incidence rates from both the National Cancer Institute – Surveillance, Epidemiology and End Results (NCI-SEER) Program, and the Canadian Cancer Society/Statistics Canada, are presented in *Table 1*.

Table 1. Age-standardized lung cancer incidence rates (per 100,000 person-years)

Source	Country	Year	Incidence (per 100,000 py)	Incidence in Males (per 100,00 py)	Incidence in Females (per 100,000 py)
NCI- SEER* (Howlader et al., 2011)	United States	2007-2011	60.0	72.1	51.1
Canadian Cancer Society† (***Statistics Canada, 2015)***	Canada	2015	51.9	57.6	47.5

Abbreviations:
py = person-years
NCI-SEER = National Cancer Institute – Surveillance, Epidemiology and End Results
****NCI-SEER Standard Population: 2000 US Population, based off single-year ages in 2000 US Census***
†Canadian Cancer Society Standard Population: 1991 Canadian Population, based off five-year age groups

Lung cancer is the second most common neoplasm in both men and women in the United States, trailing only prostate cancer in men and breast cancer in women (Howlader et al., 2011). The American Cancer Society projects 224,210 new cases of lung cancer in the United States in 2014, and an estimated 159,260 deaths from the disease; higher than the combined mortality from colon, breast and prostate cancers (American Cancer Society, 2014). While the second most common form of cancer, lung carcinoma accounts for the largest proportion of cancer deaths (American Cancer Society, 2014).

It has been estimated that between 10 and 15% of all lung cancers occur in never-smokers (Samet et al., 2009). Even using a conservative estimate of 10% of cases and the American Cancer Society's 2014 projection (American Cancer Society, 2014), this would result in approximately 22,000 new cases of lung cancer occurring in American never-smokers. There has been a great deal of variability in incidence rates for LCINS, as shown by a 2007 study by Wakelee *et al.* (2007). Amongst different cohorts, incidence

rates ranged from 14.4 to 20.8 per 100,000 person-years in women, and from 4.8 to 13.7 per 100,000 person-years in men (Wakelee et al., 2007). The cause of these discrepancies in differing cohorts can be attributed to a number of factors, including different timeframes, as well as differences in a number of different suspected etiologic factors (Wakelee et al., 2007). As is the case with most LCINS literature, the authors conclude that more research must be done on the etiologic factors associated with the disease. Nonetheless, when taken as a separate disease LCINS would rank as the 7th leading cause of cancer-related death worldwide (Samet et al., 2009). Incidence rates in the United States are similar to myeloma in men and cervical or thyroid cancer in women (Wakelee et al., 2007).

2.3 Risk Factors

In order to develop a risk prediction model for LCINS, it is important to first gain an *a priori* understanding of which factors are suspected to contribute to the etiology of the disease. While current active smoking, as well as former smoking, is an overwhelmingly strong predictor of lung cancer (10-20x increased risk in current smokers, 7.5x increased in former smokers, both compared to never-smokers) (Clément-Duchêne et al., 2010), a number of different potential demographic, lifestyle and other risk factors have been investigated with regards to their role in LCINS etiology (Samet et al., 2009). These different suspected risk factors will be discussed in the forthcoming section, in order to provide a basis for their potential inclusion as predictors in the model building stage of this thesis.

2.3.1 Modifiable Risk Factors

A number of the suspected risk factors for LCINS can be categorized as attributable to our everyday lives, or those that can be more readily changed. Henceforth in this thesis, they will be known as "modifiable risk factors", as they are pertinent to an individual's day-to-day activities and not predetermined in some way. For LCINS, these modifiable risk factors include: SHS exposure, occupational exposures, environmental exposures, BMI, hormone replacement therapy and dietary factors (Couraud, Zalcman, Milleron, Morin, & Souquet, 2012; McCarthy et al., 2012; Samet et al., 2009). Each of these risk factors, and their contributions to LCINS risk as outlined in literature, will be discussed in this upcoming section.

2.3.1.1 Secondhand Smoke

In 1986, the United States Surgeon General's report identified SHS as a cause of lung cancer in lifetime never-smoking adults (United States Department of Health and Human Services (USDHHS), 2006). In the 2006 report, which synthesizes results from a number of separate studies, it was concluded that there is sufficient evidence to suggest that SHS contains the same carcinogenic compounds as active tobacco smoking, and in concentrations great enough to cause lung cancer in lifetime never-smokers who were exposed (USDHHS, 2006).

Results from the meta-analysis conducted as part of this report examined both the at-home and workplace effects of SHS on non-smoking individuals. Upon pooling data from a number of case-control and cohort epidemiological studies investigating spousal household smoking, a range of relative risk's (RR's) was 1.20-1.29, indicating a

significant risk associated with being a non-smoker living with an active-smoking spouse, when compared to a non-smoking spouse (USDHHS, 2006). The effect size remained consistent between men and women, and across a number of different study populations and geographic areas. Similarly, for studies examining workplace exposure across a wide-range of populations, RR's varied from 1.12-1.32 when comparing non-smokers in a workplace with SHS exposure, to those without. There was a noted dose-response relationship as well, corresponding to the intensity and duration of exposure, with a 3-times greater risk experienced by those with the highest levels of exposure (>4 hours per day vs. no exposure) (USDHHS, 2006). The abundance of epidemiological evidence, coupled with the known biological carcinogenic characteristics of mainstream smoke (which composes approximately 85% of SHS), make SHS a strong risk factor for LCINS (Besaratinia & Pfeifer, 2008).

2.3.1.2 Radon Gas

Radon gas is produced from uranium present in rocks and soil, which decays into active products capable of attaching to atmospheric particles, thus allowing it to be inhaled. Upon inhalation, it emits alpha radiation which damages epithelial cell DNA (Sun et al., 2007; US National Research Council, 1999), and is considered carcinogenic. Inhaled radon gas is considered the main source of radiation exposure for the general population, as it is ubiquitous in outdoor air in low concentrations while accumulating in higher concentrations in some indoor spaces (Lagarde et al., 2001; Sun et al., 2007). Long considered a major source of lung cancer amongst underground uranium miners, more recently evidence has arisen that residential radon levels may be present in high enough

concentrations to provide a significant risk (Krewski et al., 2005; US National Research Council, 1999).

The Committee on the Biological Effects of Ionizing Radiation (BEIR) developed a model to determine the risk of residential radon concentrations, extrapolated from data obtained from studies on uranium miners coupled with typical indoor exposure levels (US National Research Council, 1999). They determined lifetime relative risks (LRRs) ranging from 1.19 to 7.01 from the lowest to highest exposure categories, indicating an apparent dose-response relationship. Furthermore, they suggested that in the United States an estimated 2,900 LCINS deaths per year can be directly attributable to residential radon exposure (Sun et al., 2007; US National Research Council, 1999). Finally, the American Cancer Society estimated that approximately 26% of all LCINS deaths are attributable to environmental radon (American Cancer Society, 2006), further indicating its role as an important risk factor.

2.3.1.3 Indoor Air Pollution

The use of biomass in cooking and household heating has been identified as a potential risk factor for LCINS due to the carcinogenic nature of its by-products (Yang, 2011). Coal smoke contains polycyclic aromatic hydrocarbons (PAH), which have known carcinogenic properties (Yang, 2011). Zhou and Christiani (2011) estimated that 13% and 17% of all lung cancers in Chinese men and women respectively were attributable to indoor air pollution, primarily coal burning and cooking fumes (Zhou & Christiani, 2011). The use of biomass for heating is especially prevalent in the developing world, with Reid *et al.* reporting that 74% of Indian, 67% of Nigerian and 81% of Kenyan households utilizing this method (Reid et al., 2012). This is a large population potentially

exposed to a risk factor for LCINS, which could lead to an increased incidence in the developing world. Many studies, such as the 2007 meta-analysis by Sun *et al.* have indicated a high relative risk for exposure to cooking fumes (approximately 2.1), and for indoor biomass burning (approximately 2.66) (Sun et al., 2007). Most studies regarding indoor air pollution are conducted on East Asian populations, and relatively fewer exist for Europe and North America. However, one Canadian case-control study indicated an OR of 2.5 for women exposed to indoor cooking fumes compared to those who were not exposed (Ramanakumar, Parent, & Siemiatycki, 2007), and no increased risk for men. The association of lung cancer and cooking fume exposure in women is consistent across many studies, as they traditionally spend more time indoors and near cooking and heating sources than men (International Agency for Research on Cancer, 2012; Ramanakumar et al., 2007).

2.3.1.4 Workplace & Occupational Environmental Hazards

Two of the predominant occupational and workplace hazards (SHS and radon) have already been discussed, so the forthcoming section looks at additional exposures suspected to be risk factors for LCINS. A third, and one that has been the subject of much discussion, is the occupational exposure to asbestos. While believed to be synergistic with cigarette smoking (Berman & Crump, 2008), other epidemiological studies suggest that exposure to occupational asbestos results in an increased lung cancer risk even among never-smokers. In their 2012 study, Villeneuve *et al.* determined an odds-ratio for men ever exposed to asbestos of 1.28, when compared to those who were never exposed. This was done while controlling for cigarette smoking, SHS and other occupational risk factors (Villeneuve, Parent, Harris, & Johnson, 2012). This study further supported the

evidence of a dose-response relationship, with the highest risk group corresponding to those with the greatest exposure (Villeneuve et al., 2012). While there is some variation in literature, small sample sizes, and limited information on specific occupations, there is generally enough information to conclude that occupational asbestos exposure serves as a risk factor for LCINS (Samet et al., 2009; Villeneuve et al., 2012).

There are a number of other occupational risk factors which have demonstrated an increased risk of lung cancer in literature, such as: silica (relative risk ranging from 1.6-2.2), arsenic (odds ratio (OR) ranging from 2.6-5.1 compared to unexposed) and various other organic solvents and non-ferrous metal fumes (Samet et al., 2009). However, McCarthy *et al.* describe these exposures, and their accompanying lung diseases, as "rare" (prevalence <1/2,000) therefore do not provide a large attributable risk to the population (McCarthy et al., 2012)

2.3.1.5 Hormone Replacement Therapy

There is a great deal of controversy in the literature about the associated risk of lung cancer in women taking hormone replacement therapy (HRT). HRT is an effective treatment for postmenopausal symptoms in menopausal females (Yao, Gu, Zhu, Yuan, & Song, 2013). In 1994, Taioli and Wylander, in their hospital-based case-control study, determined that estrogen replacement therapy (ERT), a common form of HRT, was significantly associated with adenocarcinoma (OR=1.7, 95% confidence interval 1.0-2.5, compared to those without ERT). When looking at only never-smoking women, there was found to be no significant relationship between ERT and lung cancer (OR=1.0, 95% confidence interval 0.3-3.8).

In a 2010 study by Baik *et al.*, which utilized data for 107,171 women from the Nurses' Health Study, it was concluded that there was no increased risk of lung cancer associated with the use of post-menopausal hormones (Baik, Strauss, Speizer, & Feskanich, 2010). This population was predominantly current or former smokers, and thus provides differing findings to the aforementioned Taioli and Wylander study. While the Baik *et al.* study determined no increased risk of lung cancer in women with HRT, they did identify changes in histological type associated with use of this therapy, with a shift towards increased prevalence of adenocarcinoma (Baik et al., 2010).

A 2013 meta-analysis by Yao *et al.,* which examined results for 656,403 subjects from 25 epidemiological studies spanning 24 years, found evidence of a protective effect for HRT among post-menopausal women (Yao et al., 2013). Across all the studies, an OR of 0.91 demonstrated a significant reduction in risk for women receiving HRT compared to those who did not (Yao et al., 2013). The reduction in risk was even greater in never-smoking women, with an OR of 0.86 when compared to never-smoking women without HRT (Yao et al., 2013). Given the conflicting information in literature about the risk of lung cancer among women who underwent HRT, variables pertaining to occurrence or duration of this therapy will be examined for their role as potential predictors of LCINS.

2.3.1.6 Body Mass Index

Body mass index (BMI), corresponding to height and weight in units of kg/m^2, is commonly used in literature to distinguish individuals as underweight (BMI of less than $18.5kg/m^2$), normal weight (BMI of $18.5\text{-}24.9kg/m^2$), overweight (BMI of $25\text{-}29.9kg/m^2$), or obese (BMI of at least $30kg/m^2$) (Dela Cruz, Tanoue, & Matthay, 2011; Renehan, Tyson, Egger, Heller, & Zwahlen, 2008). While increasing BMI is associated with

increased risk for many forms of cancer, this has not been the case for lung cancer. In a 2008 meta-analysis, Renehan *et al.* examined the risk of a 5kg/m^2 increase in BMI on various forms of cancer. Of the forms of cancer looked at (16 types in men, 19 in women), lung cancer was one of the few to have a protective relationship with increasing BMI; the others being esophageal squamous in both genders, and premenopausal breast cancer in women (Renehan et al., 2008). In the 13 studies- totalling 2,649,395 participants- included in the meta-analysis, a lung cancer-specific relative risk of 0.76 in men and 0.80 in women demonstrated a significant protective effect associated with 5kg/m^2 increases in BMI (Renehan et al., 2008).

One likely cause of this protective association between high BMI and lung cancer risk is confounding due to smoking; that is, smokers typically have lower BMI and significantly higher lung cancer risks (Renehan et al., 2008). When separated by smoking status, the RR for a 5kg/m^2 increase in BMI rose to 0.91 for never-smokers, and no longer demonstrated a significant protective effect (Renehan et al., 2008). In a 2007 study by Kabat *et al.*, which stratified by smoking status, an increase in BMI was associated with an increased hazard ratio never-smokers. However small sample sizes led to wider confidence intervals and thus insignificant results for all BMI groups (Kabat, Miller, & Rohan, 2007).

2.3.1.7 Diet

A number of studies have been conducted which investigate the effects of different foods, vitamins and minerals on the risk of lung cancer among both smokers and non-smokers. While the dietary intakes looked at have been variable, the most consistent protective effects seen in literature are with high vegetable and fruit intake (De Stefani et

al., 1999; Nyberg, Argenius, Svartengren, Svensson, & Pershagen, 1998; Voorrips et al., 2000; Willett & Trichopoulos, 1996). In a 1999 case-control study, which investigated the potential protective effects of antioxidants on lung cancer risk, De Stefani *et al.* determined strong inverse relationships associated with carotenoids, Vitamins C and E, glutathione and flavonoids. Of particular note were carrots, spinach, and orange consumption, all of which were deemed to be significantly strongly protective (ORs=0.55, 0.69 and 0.54 when comparing highest versus lowest tertile of carrot, spinach and orange consumption respectively) (De Stefani et al., 1999). It has been demonstrated that there is a dose-response relationship between carrots and green, leafy vegetable intake and reduced risk of lung cancer. However, it may be difficult to discern this relationship from the "healthy lifestyle" effect; individuals who consume high amounts of green, leafy vegetables would typically partake in other protective or low-risk activities (Nyberg et al., 1998).

Long thought to be one of the main protective dietary factors, β-carotene was shown to have harmful effects in the β-Carotene and Retinol Efficacy Trial (CARET) (Omenn et al., 1996). This study, which included an active intervention of 500mg β-carotene in addition to retinol, was stopped 21 months prior to intended completion due to significantly higher rates of lung cancer incidence and mortality among the intervention group compared to the control group (28% increase in lung cancer, 17% increase in mortality) (Omenn et al., 1996). In a follow-up study, carried through December 2001, it was found that lung cancer incidence was lower in the group that received the intervention than it was during the active-intervention period (RR=1.08 post-intervention versus 1.28 during intervention, comparing intervention to placebo),

indicating a reduction of effects upon cessation of β-carotene/retinol supplementation (Goodman et al., 2004). The effects of β-carotene on lung cancer have been speculated to be strengthened by an interaction with smoking, and likely caused by the inhibition of apoptosis in pre-neoplastic cells (De Stefani et al., 1999; Goodman et al., 2004).

2.3.2 Non-modifiable Risk Factors

As with many other diseases, there are a number of risk factors for LCINS that are beyond the scope of day-to-day activities of the individual. Unlike the modifiable predictors, these – henceforth known as "non-modifiable risk factors" – are less apt to be changed or influenced by a single intervention. They are more intrinsic to the individual. Among the non-modifiable risk factors most established for LCINS include age, gender, race/ethnicity, socioeconomic status, family history of cancer and a previous history of lung disease (Dela Cruz et al., 2011; McCarthy et al., 2012; Samet et al., 2009). Each of these risk factors will be discussed in the upcoming section.

2.3.2.1 Age

Described by McCarthy *et al.* as "arguably the most important risk factor among never-smokers" (McCarthy et al., 2012), the influence of age on lung cancer risk has been observed consistently. While considered a "sufficiently rare" disease amongst individuals under the age of 40 (Samet et al., 2009), lung cancer incidence rates increase greatly as a person ages, and this trend has been demonstrated amongst populations of current, former and never-smokers (Meza, Hazelton, Colditz, & Moolgavkar, 2008). The trend of greatly increasing lung cancer incidence rates amongst older populations has been observed many times, amongst many studies and reviews (McCarthy et al., 2012; Meza et al.,

2008; Samet et al., 2009; Thun et al., 2006) and is the most consistent predictor of LCINS risk. As such, its inclusion in predictive model building is necessary.

2.3.2.2 Gender

When looking at the relationship between gender and LCINS, it important to remember that the never-smoking cohort is one that is dominated by women, particularly amongst older age groups. In a 2004 study by Nordquist *et al.,* they observed 78% of the never-smoking cohort to be female, compared to just 54% in the smoking cohort (Nordquist et al., 2004). For this reason, it is important to interpret findings regarding LCINS gender differences with caution, as women represent a much larger at-risk population (Samet et al., 2009). Despite this, a 2007 meta-analysis by Wakelee *et al.,* analysing the results of 6 large cohort studies, identified an increased incidence of LCINS among women than men (range of 4.8-13.7/100,000 person-years for men versus 14.4-20.8/100,000 person-years for women) (Wakelee et al., 2007).

Although there is a higher incidence of LCINS among women than men, never-smoking women have demonstrated better survival than men (age-standardized mortality rate 10.5/100,000 person-years among never-smoking men versus 8.9/100,000 person-years among never-smoking women) (Thun et al., 2008). This aligns with the NCI-SEER data which identifies 5-year survival of 14.4% for men and 19.6% for women amongst all individuals with lung cancer relative to similar individuals without lung cancer (Howlader et al., 2011), which might suggest possible biological differences in lung cancer between genders. EGFR mutations are more prevalent in lung cancer in women, however this difference diminishes when adjusting for smoking status (Sagerup, Småstuen, Johannesen, Helland, & Brustugun, 2010). Despite this evidence of some

gender-based difference for LCINS and lung cancer in general, previous predictive models have found no significant effect of including a gender variable, indicating that it may in fact not be an important predictor (Tammemagi et al., 2011).

2.3.2.3 Race/Ethnicity

As was the case with gender, careful consideration must be made when evaluating race or ethnicity as a predictor for LCINS. As has been noted in literature, it can be difficult to discern where risk attributable to a biologic race ends and risk attributable to cultural habits or other factors associated with that race begin (e.g. dietary habits, SES, etc.) (Zhou & Christiani, 2011). However, there have been a number of consistent trends noted regarding the relationship between race and LCINS, which warrant consideration as potential predictors.

In a 2008 meta-analysis, Thun *et al.* investigated the relationship between race and LCINS among 13 cohort and 22 cancer registry studies (Thun et al., 2008). When comparing to individuals of European descent, significant age-standardized effect estimates were seen with Asian men (RR=1.96), Asian women (RR=1.69), African American women (RR=1.34), while a non-significant estimate was seen for African American men (RR=1.33, 95% CI=0.9-2.1) (Thun et al., 2008). However, the same study also noted that lung cancer death rates for Asians living in the United States were more similar to those of European descent than Asians living in Korea or Japan (Thun et al., 2008). This suggests that it is lifestyle factors associated with a certain race, and not the race itself, that may be responsible for the increased risk of LCINS in some populations. While having black race/ethnicity has been associated with a significant odds ratio in risk prediction models (OR=1.48), and lung cancer is shown to occur more frequently in

African American men (Haiman et al., 2006), incorporating "black men" versus other gender-race groups has not been shown to improve predictive model discrimination. This observation may be due to a modest number of black men in the PLCO sample, and a population that may not be representative of black men in general (Tammemagi et al., 2011, 2013). The nature of this relationship further indicates that the association between race and LCINS, or lung cancer in general, is more due to the sum of different lifestyle factors than characteristics inherent to a specified race. Despite this, race may still be useful when included in a predictive model.

2.3.2.4 Socioeconomic Status

Socioeconomic status - generally estimated as education level, income, occupation or living conditions (housing, environment) – represents an important contributing factor for a number of diseases, including LCINS. Ward *et al.* (2004) described socioeconomic factors as being "more important than biological factors" when describing their roles in cancer causation, going so far as to call poverty a carcinogen (Ward et al., 2004). This is due to the all-encompassing nature of SES. It influences so many areas of life: physical activity, occupational exposure, diet and access to proper health care, to name a few (Alberg, Brock, Ford, Samet, & Spivack, 2013; Ward et al., 2004).

In a 2009 meta-analysis of 64 studies, Sidorchuk *et al.* broke SES into educational attainment, occupational categories and income level. Individuals in this meta-analysis were categorized by Socioeconomic Position (SEP), a multidimensional measurement including any of: educational attainment, occupation, income level and other social constructs that affect health in different ways (Sidorchuk et al., 2009). In this study, they

determined that while there is a strong association seen with a decrease in any of the three factors, and an increase in lung cancer risk, the most pronounced was seen with education attainment (Sidorchuk et al., 2009). There was a 65% increase in lung cancer risk when comparing the lowest to highest educational category, while a 33% increased risk was seen comparing the lowest to highest occupational category (of the three-level SEP categorical variables) (Sidorchuk et al., 2009). These effects remained consistent in both smoking-adjusted and unadjusted studies. In previous predictive models, education levels have been used as a surrogate for SES – with lower education associated with increased risk – and thus there is reason for its inclusion in data analysis and model building (Tammemagi et al., 2011).

2.3.2.5 Previous History of Lung Disease

A number of lung and airway diseases have been implicated as potential causative agents for lung cancer, especially LCINS. The most commonly associated with LCINS are asthma, tuberculosis (TB) and pneumonia – while chronic obstructive pulmonary disorder (COPD) and emphysema appear strong only in smokers. While originally thought to be protective, due to an aversion to smoking, more recently asthma has been identified as a risk factor for lung cancer with the inflammation and inability to remove carcinogens from the airways serving as the likely causes (Fitzpatrick, 2001; McCarthy et al., 2012; Santillan, Camargo, & Colditz, 2003). In 2003, Santillan *et al.* performed a meta-analysis looking at the occurrence of lung cancer among asthmatics, both in never-smokers and in populations adjusted for smoking status. In populations limited to never-smokers, there was a fixed-effects relative risk (RR) of 1.8 (95% CI 1.3-2.3), while in a population containing current, former and never-smokers, controlling for smoking

history, there was an RR of 1.3 (95% CI 1.3-2.2) (Santillan et al., 2003). These consistent results, across multiple studies, are suggestive of asthma as a risk factor for lung cancer, particularly in a population of never-smokers (Santillan et al., 2003).

As was previously mentioned, two other lung diseases are thought to be risk factors for LCINS – TB and pneumonia. These diseases, which are both large sources of lung inflammation, have been known to expose cells to various carcinogenic compounds if the inflammation becomes chronic rather than acute (Fitzpatrick, 2001). The effects of TB and pneumonia, as well as COPD, chronic bronchitis and emphysema, were investigated in a 2011 meta-analysis by Brenner *et al*. This study, which restricted analysis to never-smokers so as to reduce confounding from smoking, determined significant positive associations between TB and pneumonia, and non-significant associations with COPD, emphysema and chronic bronchitis (RR=1.22, 95% CI 0.97-1.53 compared to COPD/emphysema/chronic bronchitis-free) (Brenner, McLaughlin, & Hung, 2011). Among the never-smoking populations, significant RRs of 1.43 and 1.90 were identified for pneumonia and TB respectively, when compared to controls (Brenner et al., 2011). While asthma, TB and pneumonia all have separate pathologies, inflammation is a common result, and thus a likely link between the three and the development of LCINS (Brenner et al., 2011; Fitzpatrick, 2001; Santillan et al., 2003).

2.3.2.6 Genetic Mutations and Familial Aggregation

Through both family history studies, and genome association studies, there has been strong evidence of a familial contribution to the risk of lung cancer, particularly among populations of never-smokers (Rudin et al., 2009). A number of different genetic mutations have been identified by Genome Wide Association Studies (GWAS) as

potentially contributing to lung cancer risk, particularly among never-smokers (Rudin et al., 2009). Three loci in particular have shown to be associated with lung cancer: *5p15, 6p21*, and *15q25* (Brennan, Hainaut, & Boffetta, 2011). Loci *15q25* codes for three cholinergic nicotine receptors, which are known to be associated with nicotine, and thus is associated with tobacco addiction in smokers. However, inheriting at least two risk alleles on this loci is associated with an 80% increase in risk, and has been shown to be independent of smoking intensity (Brennan et al., 2011). Loci *5p15* contains two genes relevant for lung cancer, particularly the telomerase reverse transcriptase (*TERT)* gene which is considered vital to many forms of carcinogenesis (Brennan et al., 2011).

Familial aggregation of lung cancer is thought to be more than genetic heritability. It also represents the accumulation of shared exposures and habits (Matakidou, Eisen, & Houlston, 2005). In 2010, Lissowska *et al.* conducted both a large, multicentre case-control study, and a meta-analysis, investigating the relationship between family history and cancer risk. When compared to individuals with no family history, there was a significant risk (OR=1.72, 95% CI 1.56-1.88) associated with having a first-degree relative with the disease (Lissowska et al., 2010). When restricted to only never-smokers, there is a reduced but still significant risk (OR=1.4 (95% CI 1.17-1.68)), indicating it is not simply a "habitual hereditary" relationship – smoking habits passed through generations (Lissowska et al., 2010). Overall, there is sufficient evidence to indicate a familial aggregation of lung cancer, both among the general population and when restricted to never-smokers, indicating its importance as a potential predictor for LCINS.

2.4 Lung Cancer Screening

While much of the focus on improving lung cancer outcomes is associated with primary prevention – smoking cessation strategies to stop the development of the disease – proper secondary preventative screening programs can also reduce mortality (van Klaveren, 2011). Approximately three quarters of all lung cancer cases are diagnosed following the onset of symptoms, at which point the disease has often progressed to a more advanced stage (Midthun, 2013). While the five-year survival for lung cancer is around 15% (van Klaveren, 2011), this is improved when successful surgical resection is completed. Among Stage IIA/IIB patients, five-year survival with resection is approximately 50%, and this improves to approximately 70% for Stage IA/IB cancers (Midthun, 2013). The improved survival and reduced mortality underscores the importance of detecting cancers at an earlier stage, where resection with curative intent is still possible, in order to improve prognosis.

Earliest forms of lung cancer screening began with chest radiography (CXR) in the 1960s-1970s, however a number of studies demonstrated no significant reduction in mortality (van Klaveren, 2011). Following the advent of low-dose computed tomography (LDCT) scanning in the early 1990s, there was an increase in lung cancer screening trials, most notably the National Lung Screening Trial (NLST), which ran from 2002 through 2009 and compared LDCT to CXR (Aberle et al., 2011). This study provided the most conclusive evidence for the efficacy of LDCT as a screening tool for lung cancer.

2.4.1 Efficacy of Low-Dose Computed Tomography Screening

The NLST was a randomized-controlled trial consisting of three rounds of annual LDCT screening in the intervention arm, and three rounds of annual CXR in the control arm (Aberle et al., 2011). The study was focused entirely on "high-risk" individuals, those with a 30+ pack-year smoking history, between the ages of 55-74 and have recently (less than 15 years) quit smoking (Aberle et al., 2011). Among the intervention arm of the study, 24.2% of all screens were positive, and 39.1% of individuals had at least one scan considered suspicious of lung cancer (with a 96.4% false-positive rate upon further diagnostic evaluation). Meanwhile, the control arm contained 6.9% positive screens, with a 94.5% false-positive rate (Aberle et al., 2011). The key finding of the NLST was that, when comparing LDCT to CXR screening, there was an approximately 20% decrease in lung cancer mortality, and a 7% decrease in all-cause mortality (Aberle et al., 2011).

The reported values for LDCT include a positive-predictive value (PPV) ranging from 2.2% to 36%, and a negative-predictive value (NPV) of approximately 99% (Humphrey et al., 2013; van Klaveren, 2011). While the low PPV indicates a high number of false-positives, the follow-up testing is often non-invasive with minimal burden on the individual – thus this is an acceptable PPV range for lung cancer screening(Aberle et al., 2011). Based on their findings, a Number Needed to Screen (NNS) – the number of individuals needed to screen to prevent one lung cancer mortality – was estimated to be 320 (Humphrey et al., 2013; van Klaveren, 2011). This is a lower number than is currently seen in breast or prostate cancer screening (van Klaveren, 2011).

2.4.2 Risks and Benefits of Low-Dose Computed Tomography

Table 2. Risks and benefits of LDCT screening

Adapted from (Wood et al., 2012)

Risks	Benefits
Futile detection of small-aggressive tumours: metastasis happens too quickly to be caught by annual screening	Decreased lung cancer mortality (20% among high-risk individuals as reported by the NLST)
Anxiety associated with test results	Improved quality of life: reduced morbidity and need for chemotherapy or radiation if cancer detected early.
Physical complications from diagnostic workup	Cost-effectiveness: $81,000 US per quality adjusted life year, below the threshold accepted for "reasonable value". More cost-effective in high-risk groups (Black et al., 2014).
High false-positive rate: 7% of false-positive cases will undergo an unnecessary invasive procedure	
Cost: with approximately 7 million "high-risk" individuals in the United States, screening costs are estimated at $2.1billion annually	
Radiation exposure: potential to cause further cancer/health complications	

2.4.3 Setting a Threshold: How Many People Do We Screen?

The NLST focused only on individuals they identified as being at the highest risk – current smokers, or those that have recently quit, between the ages of 55 and 74 years. However, it is not only this group that develops lung cancer. The key to maximizing screening effectiveness is identifying how many people we need to screen to detect the most cases while minimizing the number of false-positives. This is an important area of

healthcare research currently, as was identified by the NLST research team (Aberle et al., 2011).

While the NLST uses their aforementioned "high-risk" criteria to determine who was eligible for LDCT screening, it is likely that this method results in many missed cases of lung cancer and includes many individuals at low risk of lung cancer. Similarly, the Centres for Medicare and Medicaid Services (CMS) in the United States recently approved coverage for LDCT screening for individuals between the ages of 55 and 77 years, with a minimum 30 pack-year smoking history and either a current or recently quit (<15 years) smoker (CMS, 2015). In Canada, there are currently no recommendations regarding LDCT screening (Bell, Dickinson, & Singh, 2014). In recent years, predictive modelling – which will be discussed in detail – has gained traction as a method for identifying individuals at the highest risk using a number of different variables and modelling features. Based on a previous predictive model, using a selected probability threshold, 90% of lung cancer cases that would develop within 6 years could be detected by screening 48.7% of the highest risk smokers (Tammemagi et al., 2013). Similarly, 80% of cases could be identified by screening approximately the upper 35th percentile of highest risk individuals as identified by this model (Tammemagi et al., 2013).

In order for a screening program to be successful, it is important to carefully select who receives screening. This will identify as many cases as possible, while reducing the number of false-positives, maximizing the cost-effectiveness of screening (Aberle et al., 2011). Thus far, predictive modeling has been successful when applied to populations of smokers. However, to date, very little has been done to identify potentially high-risk never-smokers for LDCT, which is an area that will be addressed by this thesis.

2.5 Predictive Modelling

Predictive models for lung cancer have become increasingly common in recent years – used to complement clinical reasoning in modern medical decision making (Moons et al., 2012). While regression analysis can be done in a number of ways, such as for continuous outcomes, dichotomous outcomes or survival data (Vach, 2013), the focus on this section will be on modelling done for survival data– the time from the beginning of follow-up to the development of a disease, loss to follow-up, or censoring.

Predictive modeling uses a number of different covariates (predictors) to determine the probability (risk per unit time) that an individual will develop a specific outcome (Moons et al., 2012). These predictors can be any number of factors, from demographic characteristics such as age, gender or ethnicity, to clinical measurements such as blood pressure or the presence/absence of a specific biomarker (Moons et al., 2012). There are a number of different ways that candidate predictors can be selected for the final model. This is often done through a combination of *a priori* knowledge of risk factors for the outcome of interest, as well as covariates that have demonstrated good predictive performance in previous models (Royston, Moons, Altman, & Vergouwe, 2009). Two common methods of developing the final model from the set of candidate predictors are the full-model and multivariable (backward) selection approaches (Moons et al., 2012). Full-model uses all of the candidate predictors, and nothing else, and requires an extensive prior knowledge of the potential candidates for any given outcome. Multivariable selection begins with a full set of candidate predictors, and removes those that do not contribute to model performance (Moons et al., 2012). It is important not to exclude variables simply because of significance level, such as a p-value over a certain

threshold, as this can result in missing relevant predictors. (Harrell, Lee, Califf, Pryor, & Rosati, 1984; Moons et al., 2012). Conversely, choosing too many predictors (based on a high p-value), can result in model over-fitting (Moons et al., 2012). Over-fitting is a phenomenon that occurs in predictive models when the set of predictors becomes too specified to a given data sample (Royston et al., 2009). Predictors should be selected based on their contribution to predictive performance, not statistical significance. It is important to limit the number of predictors included in a model to allow for reproducibility on a new sample (Harrell et al., 1984), and thus a general rule of thumb is to have no more than one predictor for, at minimum, every 10 outcome events in the development sample (Moons et al., 2012). However, this number is not a concrete rule, and there has been shown to be limited risk associated with using between five and 16 events per variable – especially in larger data sets (Vittinghoff & McCulloch, 2007).

Predictive performance of a model is assessed primarily through two ways: calibration and discrimination. Calibration is essentially how closely the model estimated probabilities agree with the observed probabilities (Moons et al., 2012). Discrimination relates to the models ability to distinguish those who experience the outcome from those who do not – correctly identifying a case from a non-case (Moons et al., 2012). Calibration and discrimination, as well as methods for evaluating them, will be discussed in sections *2.5.1* and *2.5.2*.

2.5.1 Calibration

In order for a model to potentially achieve individual-level prediction, it must achieve a high degree of calibration. Traditionally, this has been assessed either through calibration plots of observed versus mean predicted probabilities or through a goodness-

of-fit statistic such as Hosmer-Lemeshow (p-value) (Moons et al., 2012). However, the Hosmer-Lemeshow statistic has shown to have limited power for assessing poor calibration (Royston et al., 2009). It has also shown significance with the addition of irrelevant predictors and at large sample sizes, falsely indicating poor calibration, and thus additional methods may be needed to accurately assess calibration.

One such statistic is the Brier score (Blattenberger & Lad, 1985; Brier, 1950; Murphy, 1972), roughly defined as: $\frac{\sum (\text{estimated probability - observed probability})^2}{\text{Number of observations}}$. This means that, should an individual have an estimated probability of 0.75, and have the desired outcome (observed probability=1), they would contribute a score of 0.0625 to the overall Brier score (Brier, 1950). The Brier score can be broken down into components which can be used to measure calibration and discrimination, or be used to evaluate overall model performance. While there is no accepted range corresponding to "good" calibration, a lower score represents better agreement between estimated and observed probabilities, and a score of 0.25 is equivalent to chance. Another method of evaluation is the mean and 90th percentile absolute error, corresponding to the difference between the observed probability and the model-estimated probability (Tammemagi et al., 2011). It is critical to have good calibration, especially around decision-making risk threshold.

2.5.2 Discrimination

Discrimination is often seen as the more important statistic when evaluating predictive models. It is argued that this is the case because without good discrimination, no amount of calibration can make a model accurate (Harrell et al., 1984). The most

common method of evaluating predictive model discrimination is the receiving operating characteristic – area under the curve (ROC-AUC or AUC). Conceptually, it is the proportion of all informative pairings in which the individual with the outcome scores the higher risk from the model (Hanley & McNeil, 1982). In Cox regression models, the analogous concordance statistic (c-statistic) is used, providing similar values as the AUC (Harrell, 2001). With both the c-statistic and AUC, a higher value is considered a greater degree of discrimination.

Hosmer & Lemeshow (2000) outlined a general rule for classifying AUCs (and by extension c-statistics) by level of discrimination. Based on this, they define a score of ≥0.9 as outstanding discrimination, between 0.8 and 0.9 as excellent discrimination and between 0.7 and 0.8 as acceptable discrimination. A score of 0.5 is no discrimination, equivalent to a coin flip (Hosmer & Lemeshow, 2000). According to Harrell (2001), an AUC of at least 0.8 is necessary if a model hopes to achieve acceptable individual level prediction.

2.5.3 Previous Lung Cancer Prediction Models

A number of different risk prediction models have been developed for lung cancer, largely in populations of exclusively or largely smokers. A few notable models will be discussed in this section.

2.5.3.1 Bach *et al.* 2003 Model

In response to previous models for breast cancer, Bach and colleagues developed the first multivariable model designed to predict lung cancer. This model was developed using the CARET study population, which enrolled both heavy current and former

smokers, as well as asbestos exposed men (Bach et al., 2003). Two separate 1-year models were developed using Cox proportional hazards regression: one for the probability of being diagnosed with lung cancer, and the other for the probability of dying from a competing cause. The models were cycled ten times to estimate 10-year risk. Predictors included in the model were age, gender, prior history of asbestos exposure, duration of smoking, average amount smoked per day, and duration of abstinence from smoking if a former smoker (Bach et al., 2003).

Internal validation was achieved through 10-fold cross-validation, while discrimination was assessed through the c-statistic and calibration through the evaluation of calibration plots (Bach et al., 2003). The calibration was determined to be excellent, while the c-statistic of 0.72 was in the range considered to be acceptable discrimination (Bach et al., 2003; Hosmer & Lemeshow, 2000). It was believed that the model is generalizable to other populations, however one main limitation was that it was restricted to only individuals over the age of 50 and with a history of smoking (Bach et al., 2003).

2.5.3.2 Spitz *et al.* 2007 Model

This model, developed by Spitz and colleagues (2007), was designed to expand on previous work by including risk factors beyond just age, gender and smoking characteristics. It was developed using data from a large, matched case-control study (Spitz et al., 2007). Cases were recruited from a single-centre, while controls were referred from a different network. Cases and controls were matched based on age, gender and smoking status (current, former, never), with separate models developed and validated for each smoking strata (Spitz et al., 2007). Candidate predictors which achieved significance (<5%) in univariate analysis were considered for multivariable

logistic regression, and were then eliminated via backward selection to create the final models (Spitz et al., 2007). Each smoking strata was split into two components: 75% for training, and 25% for validation. Calibration was determined via Hosmer-Lemeshow goodness-of-fit, while discrimination was evaluated by AUC for the validation set and concordance statistic for the cross-validation sets (Spitz et al., 2007).

In the never-smokers model, SHS and dust exposure, as well as a family history of two or more first degree relatives with cancer were significantly associated with an increased risk of lung cancer (Spitz et al., 2007). In the former- and current-smoker models, additional exposures were included, as well as no history of hay fever and all variables related to smoking intensity and duration (Spitz et al., 2007). While the models all demonstrated good calibration based on the Hosmer-Lemeshow values – this has already been described as not being the best evaluation method – and none appeared to have good discrimination based on the concordance statistics: 0.59 in never-smokers, 0.63 in current smokers and 0.65 in former smokers (Spitz et al., 2007). While one strength over previous models was the inclusion of never-smokers, there were a number of limitations as well. The data came from a case-control study, which increases susceptibility to selection and recall biases. Secondly, the study looked at only one ethnic group, and matched by age and smoking status – known risk factors for lung cancer (Spitz et al., 2007).

2.5.3.3 Cassidy *et al.* 2008 Model

Another model developed from a case-control study was the Liverpool Lung Project (LLP) model, developed by Cassidy and colleagues. This study used subjects between the ages of 20-80 years, collected between the years 1998 and 2005, and age,

gender and smoking-matched with 2 controls (Cassidy et al., 2008). As with the Spitz model, univariate analysis was first used to determine significance at a 5% level, and all covariates with significance were then included in initial multivariable logistic regression (Cassidy et al., 2008; Spitz et al., 2007). This was followed with backward stepwise selection, with the removal of covariates if they did not reach 5% significance in the multivariable model (Cassidy et al., 2008). 10-fold cross-validation was performed, and discrimination was assessed using AUC (Cassidy et al., 2008).

In the final model, the following covariates demonstrated significantly increased risks: family history of lung cancer (especially if diagnosed before the age of 60), prior pneumonia diagnosis, prior cancer other than lung, occupational asbestos exposure, and duration of smoking (Cassidy et al., 2008). The AUC was 0.71, which indicates good ability to discriminate high- from low-risk individuals for this model (Cassidy et al., 2008). The strengths of this model were that it compared well to previous models in terms of sensitivity and specificity, and included both the most important lung cancer risk factors (age, smoking) as well as a number of other suspected risk factors, however smoking was poorly used. It was a more detailed and included a wider range of predictive variables than the Spitz *et al.* (2007) model. Weaknesses of this model include those typical of a case-control derived model – high refusal rates and recall bias – as well as the need for further validation and limited external generalizability. Some of the predictors, such as asbestos exposure, require knowledge that may be too complex for use in a primary care setting. The reported AUC of 0.71 is on the low end of good discrimination, and may be inflated through the inclusion of never-smokers (Cassidy et al., 2008). Furthermore, the authors did not report any form of calibration for the LLP model.

Finally, while smoking is considered the strongest predictor of lung cancer, only one categorical predictor for smoking exposure was included – duration of exposure in 20-year intervals (Cassidy et al., 2008).

2.5.3.4 Tammemagi *et al.* 2011 Model

Using data from the PLCO randomized clinical trial, Tammemagi *et al.* (2011) developed a pair of models – one using the entire control arm, and one restricted to just current and former smokers in the control arm. A large number of predictors were considered for inclusion in the logistic regression model, with backward reduction at a significance level of 20% used to produce the final model (Tammemagi et al., 2011). Nonlinear effects of continuous variables were evaluated using restricted cubic splines. The overall model performance was evaluated using pseudo-R^2, while discrimination and calibration were assessed using AUC (or c-statistic) and Hosmer-Lemeshow statistics respectively (Tammemagi et al., 2011). Correcting for optimism was performed using bootstrapping internal validation techniques and external validation was assessed using the PLCO intervention arm (Tammemagi et al., 2011).

In the full control arm model, the following predictors had significant associations with lung cancer: age, lower education, lower BMI, family history of lung cancer, presence of COPD, CXR in past 3 years, being a current smoker, pack years smoked and smoking duration (Tammemagi et al., 2011). The AUC was 0.859, while the calibration slope was 0.987, which indicates both excellent discrimination and calibration. Upon external validation in the intervention arm, discrimination remained high (c-statistic=0.857) (Tammemagi et al., 2011). In the ever-smoker model, significant predictors were age, pack years and duration of smoking, while risk decreased with

increasing quit time (Tammemagi et al., 2011). As with the full control arm model, discrimination and calibration were both excellent (AUC=0.809, optimism-correct c-statistic=0.805 calibration slope=0.979) (Tammemagi et al., 2011). These models improved on previous prediction models through both the inclusion of new predictors and the use of nonlinear effects for evaluating continuous variables. Additionally, using prospective data allows for the estimation of incidence and absolute risk directly while avoiding the biases associated with case-control data (Tammemagi et al., 2011).

Using PLCO data, an ever-smoker model was modified to be applicable to the NLST population, which demonstrated excellent to good discrimination in the development and validation samples (AUC 0.803 and 0.797, respectively) (Tammemagi et al., 2013). This model, referred to as the PLCOm2012 compared positively to the NLST selection criteria (Aberle et al., 2011), indicating that this model may be suitable for individual-level recommendation for screening (Tammemagi et al., 2013). An analogous model, $PLCO_{all2014}$, was developed in the PLCO control arm ever- and never-smokers. This model achieved excellent discrimination in the PLCO intervention arm, but did not perform as well when limited to only never-smokers (Tammemagi et al., 2014).

CHAPTER III: METHODS

This chapter describes the methods and design of the study. It begins by discussing the overall study design of the PLCO randomized screening trial, including ethical considerations, recruitment, methods of obtaining consent as well as the randomization and screening processes, follow-up measures and methods of data collection and reporting. It then transitions to the design of this specific study, including the statistical methods chosen for model building, preliminary data analysis, variable selection, and methods for evaluating the completed model.

3.1 Source Data – PLCO Randomized Screening Trial

The design of the PLCO has been described in previous studies (Oken et al., 2011; Prorok et al., 2000). Conducted by the National Cancer Institute, this multicentre, two-arm, randomized trial was designed to examine the effectiveness of screening methods for prostate, lung, colorectal and ovarian cancers compared to standard medical care (Prorok et al., 2000). The pilot stage was initiated in 1993, with main study recruitment beginning in 1994 and carrying on until 2001 (Oken et al., 2011; Prorok et al., 2000). Additional follow-up is still ongoing, with expected completion in 2015 (Prorok et al., 2000).

3.1.1 Study Centres & Ethical Considerations

The PLCO used a multicentre approach, with ten centres scattered across the United States, with each centre responsible for recruiting between 5,000 and 30,000 individuals from their surrounding areas (Prorok et al., 2000), for an estimated total of ~37,000 males and ~37,000 females in each of the study arms. Each institution was

responsible for obtaining annual institutional board approval for the conduct of the study (Oken et al., 2011). The methods for obtaining consent from individual participants will be discussed in more detail in section *3.1.3*.

3.1.2 Recruitment

Participants were recruited from a variety of sources, all on a volunteer basis, with each study centre responsible for their individual recruiting practices (Prorok et al., 2000). Primary recruitment was achieved via mass mailing, and ethnic diversity was said to be reflective of the diversity of each individual study centre (Oken et al., 2011), with minority representation sought "in appropriate numbers" (Prorok et al., 2000).

Participants were excluded from the study if they had a history of any of the PLCO cancers, were currently undergoing treatment for cancer, had surgical removal of the entire colon, one lung or the entire prostate, had undergone recent screening tests, were outside of the 55-74 age range at the time of study entry, or were unwilling or unable to sign the consent form (National Cancer Institute, 2014; Prorok et al., 2000).

3.1.3 Consent

Patients were informed of any discomforts and risks associated with the screening procedures, the risk of falsely identifying cancer and were notified that diagnosis and treatment following screening would not necessarily extend a persons life (Prorok et al., 2000). Consent was received from each participant prior to randomization into the trial. The consent form was approved by the NCI, as well as the National Institutes of Health (NIH) – Office for Protection from Research Risks and the United States Office of

Management and Budget (Prorok et al., 2000). Each study centre also received approval from their respective institutional review board (Prorok et al., 2000).

3.1.4 Randomization and Screening

Upon entry into the study, baseline information regarding demographics, medical history, smoking history and past screening was obtained via a structured epidemiological questionnaire (Oken et al., 2011). Block randomization was utilized to randomize participants to the intervention or control arms, stratified by study centre, gender and age (Oken et al., 2011; Prorok et al., 2000). Individuals assigned to the screening arm were to receive CXR for lung cancer, in addition to screening tests for colorectal, prostate (if male), and ovarian (if female) cancers (Prorok et al., 2000). At the time of randomization, an additional dietary questionnaire (DQX) was offered to participants in the intervention arm, and 82% were completed (National Cancer Institute, 2014).

In total, 154,901 participants were randomized, with 77,445 entering the internvetion arm and 77,456 entering the control arm (Oken et al., 2011). Individuals in the screening arm received baseline and three annual rounds of CXR taken by a qualified technologist and interpreted by a radiologist (Prorok et al., 2000), while those in the control arm continued to receive standard medical care. Never-smokers randomzied post-April 1995 received only two follow-up CXR exams (Prorok et al., 2000). An examination was considered positive if it contained any of a number of abnormalities suspicious of lung cancer, as determined by a radiologist (Prorok et al., 2000).

3.1.5 Diagnostic and Therapeutic Follow-up

Individuals with positive screening results were notified, and referred to a physician of their choosing for appropriate follow-up (Prorok et al., 2000). While the PLCO protocol does not describe specifications for diagnosis and therapy, participants with a positive screening result were recommended to seek appropriate medical follow-up including diagnosis and, if required, treatment (Prorok et al., 2000). All treatment was expected to be in accordance with currenct accepted practice for stage of cancer, age, and overall medical condition of the participant (Prorok et al., 2000). Along with screening tests, all complications resulting from diagnostic or therapeutic follow-up were recorded (Prorok et al., 2000).

3.1.6 End Points

The primary endpoints for the PLCO screening trial were cause-specific mortality for each of prostate, lung, colorectal, and ovarian cancers (Prorok et al., 2000). Secondary endpoints include incident cases of cancer, stage shift, and case survival (Prorok et al., 2000). Information regarding the any-site diagnosis of cancer as well as all deaths occuring during the trial were obtained by annual study update questionnaires, and follow-up phone calls if necessary (Oken et al., 2011) while end-points were verified through linkage to the National Death Index (NDI) (Oken et al., 2011).

3.1.7 Data Recording and Follow-up

All prevalent and incident PLCO cancers, and cause-specific deaths that occur during follow-up were ascertained by an active follow-up process and supplemented through usage of cancer registries if such data were available to each individual study

centre (Prorok et al., 2000). Participant information was recorded for the following: identification number, demographic and risk factor information, randomization group, date of birth and date of entry, results of each screening test including any complications, sufficient information on any diagnostic follow-up, any PLCO cancer diagnosed during follow-up and information regarding histology and stage at diagnosis, and every death – time and cause – occurring in both trial arms (Prorok et al., 2000). In this study, all personal identifiers (date of birth, date of death) have been removed to ensure anonymity.

A total of 13 years of follow-up was conducted for each participant who was not diagnosed with cancer or lost to follow-up in both the intervention and control arms, with an estimated completion in 2015 (Prorok et al., 2000). In 2006, a supplemental questionnaire (SQX) was sent to any remaining study participants. This questionnaire collected overlapping information with what was collected at baseline, as well as new information pertaining to some occupational exposures, physical activity, history of asthma, and SHS (National Cancer Institute, 2014).

3.2 Risk Prediction Model Development and Validation

This upcoming section discusses the methods for model building from the initial stages through completion and validation. Topics include the choice of statistical methods, candidate predictor selection, data maintenance, as well as the model building and evaluation, and methods for validating the model.

3.2.1 Sample Data

Data for this study came from the aforementioned PLCO randomized screening trial, utilizing six-years of follow-up of never-smokers in both the intervention and

control arms. This was done to develop a comparable model to those by Tammemagi *et al.* (Tammemagi et al., 2013, 2014). The PLCO contains 69, 272 never-smokers, among which there were 276 cases of lung cancer (109 occurring in the six-year follow-up) – representing one of the largest never-smoker cohorts analyzed to date.

3.2.2 Statistical Methods

Since the study contains both a dichotomous outcome measure and time-to-event data, Cox proportional hazards regression was the best choice for statistical methodology for the predictive model (Cox, 1972). The outcome measure was lung cancer incidence (yes or no) at any point during the study period – screening or follow-up up to six years. Time-to-event data were the time from randomization to the incidence of lung cancer, or until the six-year cap or time to lost to follow-up in those without the disease. All preliminary variable analysis, data maintenance, model building and evaluation were performed using Stata 13 Statistical Software (StataCorp. 2013. *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP) and the *rms* package in R-statistical software (Harrell, 2014).

3.2.3 Candidate Predictors

Candidate predictors were chosen based on *a priori* knowledge of LCINS risk factors - those described in the literature review, combined with variables that have demonstrated good predictive ability in previous models. The variables preselected as candidate predictors include the following demographic risk factors: age, gender and race/ethnicity, and education level (six-level categorical) (Tammemagi et al., 2011).

Body mass index, measured in kg/m^2, was considered as a potential protective predictor for LCINS.

History of COPD, chronic bronchitis, and emphysema were all considered as they pertain to previous history of lung diseases. A number of variables were considered regarding family history – lung cancer incidence in father, mother, sibling or child – as well as the age at which these individuals were diagnosed. Receiving a CXR in the three years prior to baseline was also considered, as it has demonstrated an increased risk of lung cancer in previous models (Tammemagi et al., 2011). Three variables were investigated to address the conflicting literature regarding the association of female hormones and LCINS: is the individual currently taking female hormones (yes vs. no), have they ever taken them (yes vs. no) as well as the age of onset of menopause (<40 vs. 40+). Variables pertaining to dietary intake – particularly fruits and vegetables – and physical activity were investigated. Finally, SHS exposure was evaluated using three variables from the supplemental questionnaire: household smoking prior to age 18 years, household smoking after the age of 18 and indoor workplace smoking, all of which are dichotomous (yes vs. no) variables. Given the limitations of previous models in never-smokers, novel predictors were investigated to potentially improve predictive performance. Limited exploratory analysis was conducted for some suspected – but less established – risk factors for lung cancer, such as past ibuprofen use (Endo, Yano, Okumura, & Kido, 2014).

A number of predictors could not be included, despite known or suspected association with LCINS, due to limitations in the comprehensiveness of the PLCO data

collection. Risk factors that could not be considered as candidates include radon exposure, asbestos and other occupational exposures, and indoor air pollution.

3.2.4 Data Preparation and Maintenance

Prior to regression modelling, individual predictors were separately evaluated for missing data, as well as potential outliers. Outliers were investigated using boxplots, with outliers defined as values 3-times the interquartile range (25^{th}-75^{th} percentile) above the 75^{th} percentile or below the 25^{th} percentile (Steyerberg, 2009). However, the PLCO utilized Information Management Services to ensure clean data, and thus minimal cleaning was required. Measures of central tendency were calculated to determine the spread of continuous variables. To examine differences in descriptive statistics between individuals with lung cancer and those without, Wilcoxon rank-sum test was used for non-normally distributed continuous variables and Pearson's chi-square test or Fishers exact test were used for categorical variables.

3.2.4.1 Multiple Imputation

Variables with a large amount of missing data, such as those from the Dietary and Supplemental questionnaires, were imputed to maximize the number of observations included in the final model. Data can be missing in one of three ways: missing completely at random, missing at random, or missing not at random (White, Royston, & Wood, 2009). Missing data in the PLCO data set was determined to be missing at random through the use of the *patterns* function in Stata as well as consideration for the cause of the missingness. Data for these variables was missing for a reason (participants did not complete the questionnaires), however this was not related to the variables themselves.

Missing at random allowed the missing-ness to be handled through multiple imputation (White et al., 2009). Due to the many types of variables being imputed, multiple imputation by chained equations (MICE) was chosen as the optimal method. MICE allows for simultaneous handling of continuous and categorical variables, and even non-normally distributed continuous data (White et al., 2009).

Ten imputations were chosen, as it provided a large enough number to avoid skewed estimates, while still allowing completion in a reasonable timeframe (White et al., 2009). Previous history of cancer, CXR in the past three years, ibuprofen usage, and the SHS variables (as a child, workplace, and as an adult) were imputed as logit variables. Family history of lung cancer, education level, and marital status were imputed as multilevel-logit variables. Drinks per week and BMI were imputed using predictive mean matching, a method of imputing non-normally distributed continuous variables. Finally, age, gender and lung cancer diagnosis were used as predictors in addition to the aforementioned imputed variables, but not imputed as they did not contain missing data. The effects of variables in the final model were then combined from each imputation, as was allowed by Rubin's rules (D. B. Rubin, 1987).

3.2.5 Model Building

3.2.5.1 Handling Continuous Predictors

The continuous predictors considered for the final prediction model, age and BMI, were not categorized so as to avoid losing information unnecessarily. Non-linear effects of continuous predictors were explored using multiple fractional polynomials (MFPs), allowing for easier calculation of risk probabilities while being less prone to over-fitting

than some alternative methods. Continuous variables were centred near their approximate mean to facilitate the interpretation of results, graphing of associated risk, and preparing individual risk values.

3.2.5.2 Variable Selection

Beginning with the full set of candidate predictors, model reduction was accomplished through "intelligent" variable selection. This method removed variables based on their relative contribution to the model predictive performance – how they affected the c-statistic – rather than an arbitrary p-value cut point. Variables were only be removed from the model if they demonstrated a small or implausible effect in the model, or if their contribution to model prediction was minimal (Steyerberg, 2009). Reducing the number of variables in the model, and limiting this model to those that may be more readily available, reduced the likelihood of over-fitting to this sample and also made the model more feasible for public and clinical use (Steyerberg, 2009).

Interactions that had a plausible association with lung cancer risk, such as gender-race and gender-alcohol consumption, were evaluated using a Wald test for contribution to model performance. Collinearity was first analyzed through a correlation matrix, inputting all variables to be included in the model. Associations between categorical variables were further analysed using Pearson's chi-square analysis. Any predictors that demonstrated a strong degree of correlation were evaluated. Potential combination into a new variable, separate evaluation, or removal from the model entirely were considered depending on the nature of the variable (Steyerberg, 2009).

3.2.5.3 Assumption Checking

One assumption that must be satisfied to indicate that the model has been adequately fitted is the proportional hazards assumption – stating that the hazard function for two individuals will remain constant over time (Cox, 1972; Ng'andu, 1997). There are a number of different ways that the proportional hazards assumption can be tested, however in this study we used Schoenfeld residuals as a function of time – accomplished using the "*estat phtest"* function in Stata (Cleves, Gould, Gutierrez, & Marchenko, 2008). Assessment was accomplished through the graphical interpretation of the scaled Schoenfeld residuals plotted against time, with a slope of zero indicating that the hazards are proportional and the assumption is satisfied (Cleves et al., 2008).

3.2.6 Model Evaluation

3.2.6.1 Discrimination

The importance of discrimination as an evaluator of a models predictive performance was discussed previously in this thesis. Since this model was created using Cox survival analysis, the c-statistic – analogous for the AUC – was used to evaluate discrimination (Harrell, 2001). The c-statistic for the full model was estimated as a mean of each of the 10 multiple imputation models. A comparable logistic model was created for comparison purposes. Discrimination was categorized using Hosmer-Lemeshow AUC cut-points (Hosmer & Lemeshow, 2000), with a score of 0.7 or greater corresponding to good-to-excellent discrimination, and between 0.6 and 0.7 corresponding to adequate discrimination. The aforementioned Tammemagi *et al.* PLCOall2014 model was tested only on the never-smoking cohort, and performed with an AUC of 0.662 (Tammemagi et

al., 2014). This AUC will provide a benchmark for which to improve upon in the current study.

3.2.6.2 Calibration

The common methods for evaluating calibration were also discussed previously in this thesis. This study proposed to use three methods: the Spiegelhalter's z-statistic component of the Brier score (Brier, 1950), visual interpretation using calibration plots (predicted versus actual probability) (Steyerberg, 2009), and the mean and 90th percentile absolute error – which has been used in previous predictive studies (Tammemagi et al., 2011, 2014). There is no accepted range for a "good" Brier score, however a lower score indicates better calibration. Calibration plots were evaluated against a slope of 1 (perfect calibration) while the mean and 90th percentile errors were compared to the Tammemagi *et al.* (Tammemagi et al., 2014) values for the never-smoker cohort (mean=0.0002, 90th percentile=0.0003).

3.2.6.3 Internal Validation

In order to maximize the data available for model building, the full control and intervention arms were used for training. Thus, there was no portion of the data set aside for validation as had previously been commonplace in predictive studies (Moons et al., 2012). Instead, bootstrapping validation techniques were used. Bootstrapping involves the repeated sampling with replacement of the original study data, creating numerous bootstrap samples with which to estimate the overall model performance corrected for over-fitting (Steyerberg, 2009). In this analysis, 2000-times bootstrap resamplings were performed using R-statistical software, specifically Harrell's *rms* package (Harrell, 2014).

3.2.7 Predicted Probabilities

In order to present the results from the model in a meaningful and clinically relevant way, predicted probabilities were constructed for all individuals in the sample for whom data were available. These are often presented as the probability of developing lung cancer within a set time frame – one year, five years, ten years, etc., and provide a useful and easily understood tool for clinicians (Moons et al., 2012). In Cox models, the predicted probabilities are calculated through the individual variable values – an individual's age, for example – multiplied by the coefficient determined by the model. The sum of all of the predictor values provides the raw risk score, which must then be translated into an overall probability by incorporating the baseline survivor function – the probability of an individual with values of x for all the predictors developing the disease at a given point in time – x being the mean or proportion for each variable (Vach, 2013; Woodward, 2013). For this model, predicted probabilities were calculated for six-year lung cancer incidence. Six-year probabilities were calculated to compare to the Tammemagi *et al.* PLCO models (Tammemagi et al., 2011, 2014).

CHAPTER IV: RESULTS

4.1 Population Characteristics

Characteristics of the 69,272 never-smokers in the PLCO control and screening arms are described in *Table 3.* Information for this table consists of baseline characteristics for individuals, in addition to responses from the Dietary and Supplemental questionnaires for available participants. The participants were divided into two groups, those who never experienced a lung cancer diagnosis at any point during follow-up (n=68,996), and those that did (n=276). The overall incidence of lung cancer amongst PLCO never-smokers was 35.4 cases/100,000py over the complete follow-up (median = 12.5 years). Differences between groups were evaluated using the Wilcoxon rank-sum test for non-normally distributed continuous variables (age, BMI), chi-square test for categorical variables, and Fishers exact test for categorical variables with cells containing five or fewer individuals. There were significant differences seen between groups for age, BMI, personal history of cancer, family history of lung cancer, and CXR in the past three years. No outliers were detected or removed, a product of the rigorous follow-up and data management procedures of the PLCO data set.

Univariate hazard ratios and 95% confidence intervals are presented for each variable, estimating the six-year risk of developing lung cancer. This univariate analysis was also used to guide variable selection for model building, which is discussed in a later section. The strongest significant effect was seen with previous history of cancer (HR=2.57, 95% CI 1.25-5.28). A larger effect was seen with SHS exposure as an adult, however this was non-significant (HR=3.81, 95% CI 0.95-15.23).

Table 3. Characteristics of PLCO never-smoker population (N=69 272)

Variable	No lung cancer (n= 68 996)	Lung cancer (n= 276)	P	Number of Missing	Univariate Hazard Ratio (95% CI; p-value)
Sociodemographic					
Age, mean (SD), years	62.86 (5.43)	65.56 (5.21)	<0.001*	0	1.10 (1.07-1.14; <0.001)
Gender, number			0.688 †	0	$1.25_{\text{female vs male}}$ (0.84-1.85; 0.276)
Female	41 183 (99.59%)	172 (0.41%)			
Male	26 813 (99.61%)	104 (0.39%)			
Race/ethnicity, number			0.798‡	23	
White	61 067 (99.60%)	248 (0.40%)			1.0
Black	3 118 (99.71%)	9 (0.29%)			0.39 (0.10-1.57; 0.184)
Hispanic	1 177 (99.75%)	3 (0.25%)			----
Asian	3 101 (99.55%)	14 (0.45%)			0.77 (0.28-2.09; 0.607)
American Indian	361 (99.45%)	2 (0.55%)			1.67 (2.33-11.99; 0.609)
Pacific Islander	149 (100%)	0 (0%)			----
Education, number			0.439†	177	0.89 (0.80-1.01; 0.061)
Less than HS	3 998 (99.45%)	22 (0.55%)			
HS graduate	16 505 (99.58%)	70 (0.42%)			
Post-HS training	7 867 (99.57%)	34 (0.43%)			
Some college	13 885 (99.63%)	52 (0.37%)			
College graduate	12 102 (99.68%)	39 (0.32%)			
Postgraduate	14 465 (99.61%)	56 (0.39%)			
Medical history					
Body mass index, mean (SD), kg/m^2	27.24 (4.99)	26.63 (4.78)	0.036*	1048	0.97 (0.93-1.01; 0.184)
Personal history of cancer, number			0.017†	0	2.57 (1.25-5.28; 0.010)
Absent	66 942 (99.61%)	261 (0.39%)			
Present	2 054 (99.28%)	15 (0.72%)			
Family history of lung cancer, number			0.037‡	2210	1.57 (0.97-2.54; 0.065)
Absent	60 234 (99.62%)	230 (0.38%)			
One relative	6 235 (99.52%)	30 (0.48%)			
Two or more	329 (98.80%)	4 (1.20%)			

(*continued on the following page*)

Variable	No lung cancer (n= 68 996)	Lung cancer (n= 276)	P	Number of Missing	Univariate Hazard Ratio (95% CI; p-value)
Ever diagnosed with COPD?			0.109‡	0	073 (0.23-2.30; 0.588)
No	66 404 (99.59%)	271 (0.41%)			
Yes	2 592 (99.81%)	5 (0.19%)			
Chest x-ray in the past 3 years, number			0.009†	2987	1.85 (1.21-2.83; 0.005)
None	32 246 (99.65%)	112 (0.35%)			
One	21 655 (99.60%)	87 (0.40%)			
Two or more	12 120 (99.47%)	65 (0.53%)			
Exposure history					
Regular ibuprofen use, past 12 months			0.031†	348	0.97 (0.64-1.47; 0.875)
No	49 478 (99.57%)	212 (0.43%)			
Yes	19 174 (99.69%)	60 (0.31%)			
Drinks per week, number			0.316†	14 140	0.94 (0.78-1.12; 0.457)
None	19 548 (99.65%)	69 (0.35%)			
Less than one	15 595 (99.68%)	50 (0.32%)			
Between 1 and 3	7 967 (99.61%)	31 (0.39%)			
Between 3 and 7	5 552 (99.69%)	17 (0.31%)			
Between 7 and 14	4 135 (99.73%)	11 (0.27%)			
14 or more	2 144 (99.40%)	13 (0.60%)			
Live with a smoker as an adult, number			0.297†	31 762	3.81 (0.95-15.23; 0.059)
No	33 091 (99.80%)	66 (0.20%)			
Yes	4 341 (99.72%)	12 (0.28%)			
Live with a smoker as a child, number			0.338†	31 940	0.92 (0.22-3.83; 0.854)
No	22 508 (99.78%)	50 (0.22%)			
Yes	14 748 (99.82%)	26 (0.18%)			
Work with a smoker as an adult, number			0.183†	31 793	2.34 (0.49-11.29; 0.698)
No	33 342 (99.81%)	65 (0.19%)			
Yes	4 060 (99.71%)	12 (0.29%)			

*** P-value by Wilcoxon rank-sum test**
† P-value by chi-square test
‡ P-value by Fishers exact test

4.2 Predictive Model in PLCO Never-smokers

For the final predictive model, beta-coefficients and hazard ratios were estimated using the 10-times multiple imputation (full) model, and the PLCO never-smoker population (n=68,735). Univariate analysis (*Table 3)* was used to help select candidate predictors, and final variable selection was based on a combination of statistical significance ($p<0.3$), contribution to model performance, and plausibility of the relationship between predictor and outcome. In addition to the 10-times multiple imputation model, a completed cases model (n=34,355) was created for comparison using only non-imputed data. An overview of both the multiple imputation and completed cases models is presented in *Table 4.*

In the full model, lung cancer risk increased with increasing age, lower BMI and lower education level. Furthermore, having a previous personal history of any cancer, more than one CXR in the past three years, and living with a smoker as an adult were all associated with an increased risk of lung cancer. Having one first-degree relative with a history of lung cancer was associated with an increased risk (HR=1.336, 95% CI=0.745-2.128), while having two or more first-degree relatives was associated with a further increased risk (HR=3.522, 95% CI=0.865-14.354). The low number of events in the completed cases model indicates the necessity of multiple imputations.

Table 4. Multiple Imputation and Completed Cases Cox Proportional Hazards models, six-year follow-up.

	Multiple Imputation Model (n = 68 735)	Complete Cases Model (n= 34 355)
Variable	**Hazard Ratio (95% CI; p-value)**	**Hazard Ratio (95% CI; p-value)**
Lung cancer cases, n	109	9
Age, per year	1.095 (1.057-1.135; <0.001)	1.039 (0.917-1.176; 0.549)
Body mass index, per kg/m^2	0.972 (0.932-1.013; 0.184)	0.941 (0.803-1.102; 0.451)
Education, per 1 of 6 levels change	0.935 (0.829-1.054; 0.272)	1.106 (0.719-1.699; 0.647)
Personal history of cancer, yes vs no	1.909 (0.752-4.850; 0.173)	----
Family history of lung cancer		
No relatives	Reference group	Reference group
One relative	1.336 (0.745-2.128; 0.331)	1.337 (0.164-10.906; 0.787)
Two or more relatives	3.522 (0.865-14.354; 0.079)	23.250 (2.823-193.657; 0.003)
Chest x-ray in the past 3 years, more than 1 vs 1 or fewer	1.554 (1.004-2.408; 0.048)	0.623 (0.077-5.040; 0.657)
Live with a smoker as an adult, yes vs no	1.262 (0.555-2.872; 0.567)	3.901 (0.936-16.259; 0.062)
Model Performance		
Harrell's c-statistic	0.6840 (0.6770-0.6963)*	0.7059 (0.5052-0.9066)
Optimism corrected	0.6645†	----
Area under the curve (AUC)	0.6858 (0.6358-0.7358) §	0.5696 (0.4912-0.6351) §‡
Absolute error		
Mean	0.0018‖	0.0018
90th percentile	0.0027‖	0.0032
Brier score	0.0016 ¶	0.0003 ¶
Spiegelhalters statistic (p-value)	0.9826	>0.999

***Mean of 10 imputations, range provided instead of 95% CI**
† 2000x bootstrap optimism correction using Harrell's *rms* package in R software
‡1000x bootstrap corrected for optimism
§ Calculated using Stata "comproc" function and Cox model-estimated probabilities
‖ Calculated using representative imputed dataset (Imputation #6)
¶ Brier score calculated using six-year lung cancer incidence and model-estimated six-year risk

Education level was treated as a continuous variable, due to the relatively consistent change in effect between increasing levels. Non-linear effects of continuous variables – age and BMI – were investigated using MFPs, however the relationship with lung cancer risk remained linear, and thus the variables were only centred at their approximate means (62 years and 27 kg/m^2 respectively).

Variables such as gender, race/ethnicity, lung comorbidities, alcohol consumption, amount of self-reported physical activity, SHS exposure in the workplace and as a child (<18 years of age), and regular ibuprofen and aspirin use were investigated based on suspected association, however they were rejected from the final model due to weak, non-significant or implausible effects on lung cancer risk. Interaction terms were investigated, however none contributed to model performance while also being a plausible relationship.

4.2.1 Model Evaluation

Statistics evaluating model performance are presented in *Table 4*. For the full model, a c-statistic was generated from the mean c-statistics for each of the 10 individual imputations (c=0.684, range 0.677-0.696). The Cox model performed similarly in male- and female-specific populations, with an estimated c-statistic of 0.694 (range 0.686-0.711) and 0.675 (range 0.669-0.683) respectively. Through internal validation, via 2000-times bootstrap resampling, an optimism corrected c-statistic was estimated to be 0.665. A comparable logistic model, created using the Cox model-estimated six-year risk and using six-year lung cancer incidence as the outcome, produced a ROC-AUC of 0.686 (95% CI 0.636-0.736). These c-statistics and ROC AUCs correspond to fair or adequate discrimination, falling just short of what some consider to be good discrimination (>0.7).

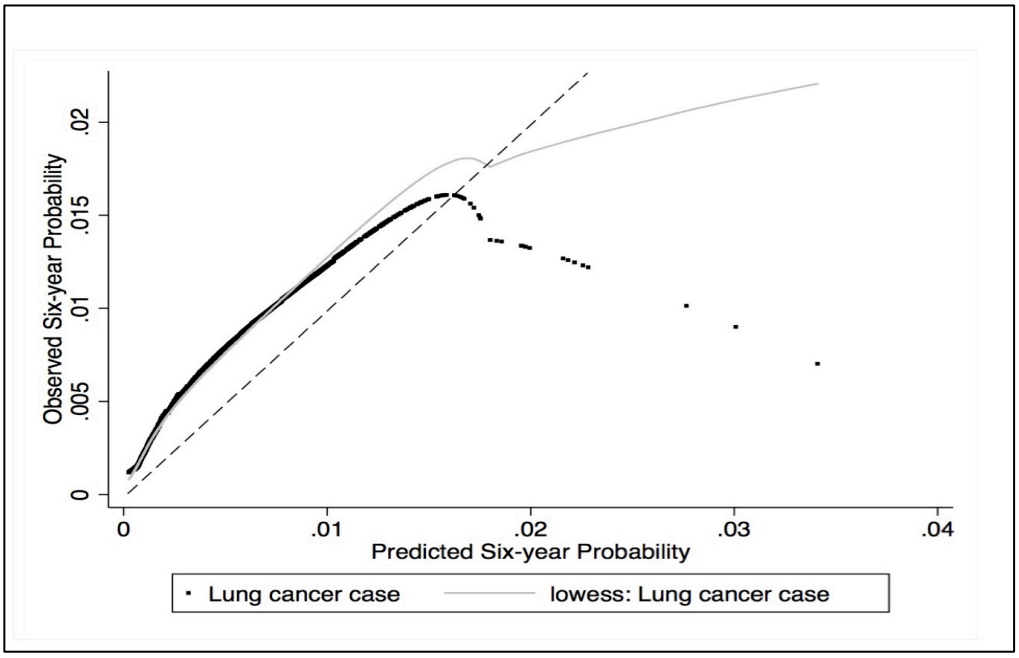

Figure 1. Calibration plot: observed vs. predicted six-year lung cancer probabilities.

Model calibration was evaluated through two methods: mean and 90th percentile absolute error, and a calibration plot (observed versus predicted lung cancer probabilities). In the full model, the mean and 90th percentile absolute errors were 0.0018 and 0.0027 respectively (*Table 4*), which compares favourably to other models (Tammemagi et al., 2011). The calibration plot, seen in *Figure 1*, displays the observed and predicted six-year lung cancer probabilities, a lowess curve plotted from these data, and a dashed-line displaying a slope of 1.0 (perfect calibration) for reference. Finally, a Brier score was calculated to evaluate overall model performance (full model Brier score = 0.0016, Spiegelhalter p-value = 0.983). The current model was compared to a close approximation of the $PLCO_{all2014}$ model, as presented in *Figure 2.* This demonstrated a modest, however non-significant improvement in model discrimination.

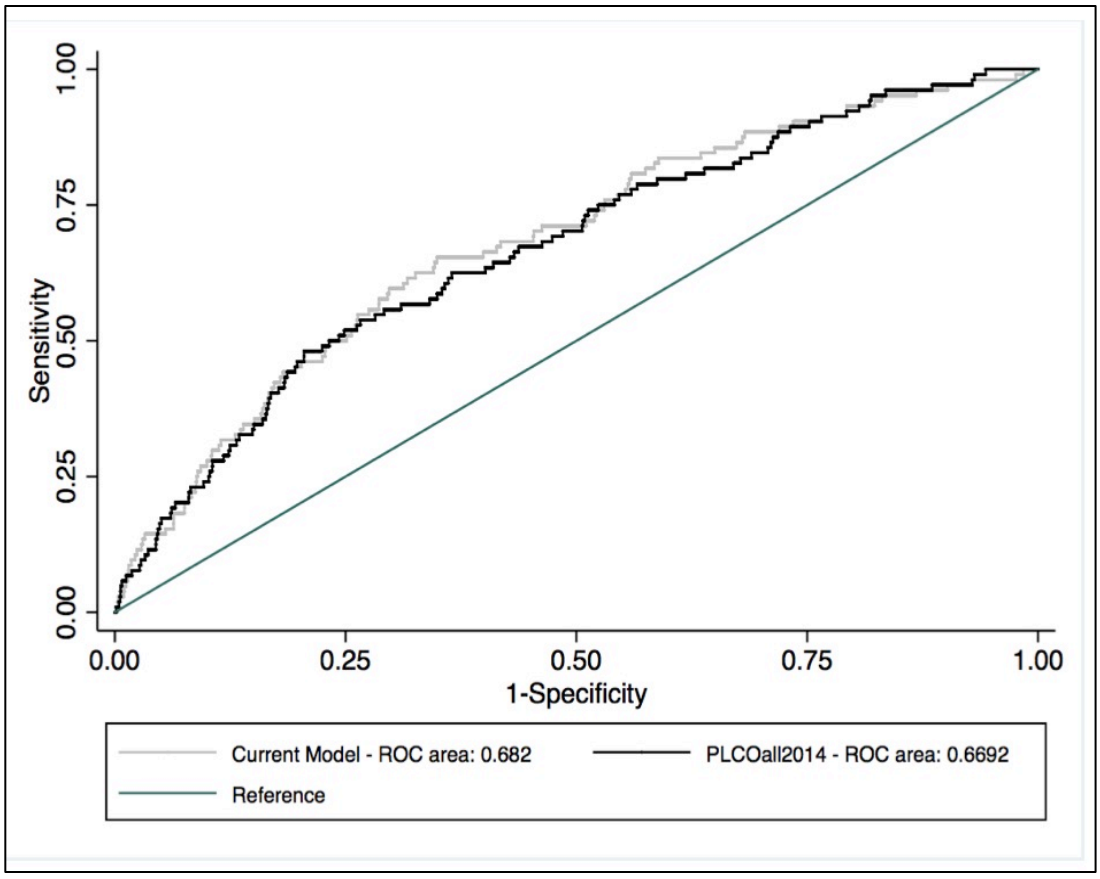

Figure 2 Comparison of the ROC-AUC for the current model and a close approximation of the $PLCO_{all2014}$ model using the PLCO never-smoker sample (n= 64 752)

4.3 Six-year Predicted Probabilities

Predicted probabilities, the model-estimated risk of developing lung cancer in six years, were calculated for each individual in the full model sample. The first component of this, the baseline survival probability, was determined to be 0.9984 from both an actuarial and Kaplan-Meier life table. The actuarial life table is presented in *Table 5*.

Table 5. Actuarial life table. Estimating year-to-year probability of remaining lung cancer-free.

Interval		Number at beginning	Number of new cases	Number lost	Survival probability
0	1	69 272	22	589	0.9997
1	2	68 661	20	473	0.9994
2	3	68 168	18	542	0.9991
3	4	67 608	18	547	0.9989
4	5	67 043	18	602	0.9986
5	6	66 423	14	639	0.9984
6	7	65 770	22	700	0.9980
7	8	65 048	32	753	0.9975
8	9	64 263	23	1 984	0.9972
9	10	62 256	21	6 150	0.9968
10	11	56 085	30	10 299	0.9962
11	12	45 756	19	9 798	0.9958
12	13	35 939	19	35 920	0.9947

The baseline survival probability, when combined with beta-coefficients (equation in *Table* 6) from the full model, produced an algebraic risk calculator for individual six-year lung cancer risk. The formula for calculating individual risk probabilities is displayed in *Table 6*.

Table 6. Algebraic six-year risk probability equation and beta-coefficients for individual predictors.

Equation Probability = $1 - (0.9984e^{(\beta 1x1 + \beta 2x2\ldots)+0.021098606)})$*	
Beta-coefficient (β)	**x-value (*x*)**
0.0910496	Age (years) - 62
-0.0283397	BMI (kg/m^2) – 27
-0.06755	Education level: 1= less than high school 2= high school graduate 3= post high school training 4= some college 5= college graduate 6= postgraduate
0.6467674	Previous history of cancer: 0= no 1= yes
0.2904569	Family history of lung cancer: 0= no relatives, or more than one 1= one relative
1.259542	Family history of lung cancer: 0= one or fewer relatives 1= two or more relatives
0.4414639	CXR in the past three years: 0= one or fewer occasion 1= more than one occasion
0.2333267	Lived with a smoker as an adult (>18): 0= no 1= yes

*** Equation adapted from *Epidemiology: Study Design and Data Analysis*, p. 898 (Woodward, 2013)**
0.021098606 represents mean or proportion of each variable multiplied by respective beta-coefficient

For the full model sample, the mean predicted risk was 0.19%, while the highest risk achieved by any individual was 3.42%. The highest risk individual in this population was a 73-year old college graduate, with a BMI of $30.8kg/m^2$, previous personal history of cancer, multiple first-degree family members with lung cancer, and multiple CXRs in the past three years. They had no response to whether they lived with a smoker as an adult. The distribution of the predicted risk values is presented in *Figure 3*. The model predicted six-year lung cancer risk, divided into approximately equal sized deciles, is

presented in *Table 7.* This table demonstrates a higher predicted risk of lung cancer in the higher risk deciles, corresponding to an increased incidence.

Table 7. Deciles of model-predicted six-year lung cancer risk with accompanying observed lung cancer probability and mean predicted risk in the same timeframe.

Decile of Predicted Risk	Number of Individuals	Number of Lung Cancer Cases in Six Years	Observed Lung Cancer Probability	Predicted Lung Cancer Probability
1	6 907	4	0.06%	0.06%
2	6 903	3	0.04%	0.08%
3	6 905	5	0.07%	0.10%
4	6 904	5	0.07%	0.11%
5	6 905	14	0.20%	0.14%
6	6 905	6	0.09%	0.16%
7	6 904	7	0.10%	0.20%
8	6 905	15	0.22%	0.25%
9	6 905	20	0.29%	0.31%
10	6 904	30	0.43%	0.52%

4.3.1 Comparison to Established Screening Criteria

The utility of the full model as a practical screening selection tool was evaluating by comparing it to the suggested risk cut-point from the $PLCO_{m2012}$ model (risk ≥1.51%) (Tammemagi et al., 2013, 2014). Applying the $PLCO_{m2012}$ criteria – with the full model derived six-year risk score - to the PLCO never-smoker population, results in 35 individuals (0.05% of PLCO never-smokers) being recommended for further screening. Of these 35 individuals, none received a lung cancer diagnosis within six years of follow-up.

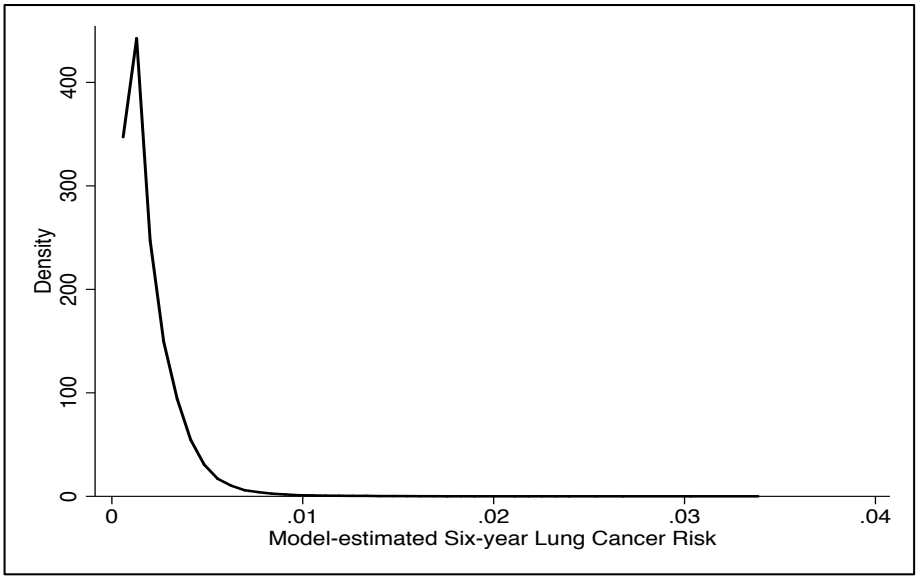

Figure 3. Distribution of model-estimated six-year lung cancer risk for full model sample. Figure created through *kdensity* function in Stata 13 displaying kernel density of model-estimated probability.

4.4. Assumption Checking

The proportional hazards assumptions was evaluated using both the *phtest* function in Stata, as well as through the examination of Schoenfeld residual plots for each individual predictor. Based on the *phtest*, every variable passed the proportionality assumption ($p>0.05$), while the full model passed the Global test ($p=0.7774$), indicating a hazard function parallel with baseline over time. Complete results from the proportional hazards test is presented in *Table 8.*

Table 8. Results of the proportional hazards test, for both individual level variables and full model.

Variable	P-value
Age (centred at 62years)	0.2470
BMI (centred at 27kg/m^2)	0.4401
Education level	0.1030
Previous history of cancer	0.7791
Family history of lung cancer	
One first degree relative	0.4147
Two or more first degree relatives	0.9068
Chest X-ray in past 3 years	0.9904
Lived with a smoker as an adult	0.8885
Global test	0.7774

To further ensure that the proportional hazards assumption was met, visual examination of each Schoenfeld residual plot was performed. Through this examination, each variable appeared to have a residual slope of approximately zero, indicating proportionality. As an example, the Schoenfeld residual plot of the variable age (centred on 62 years) appears in *Figure 4*.

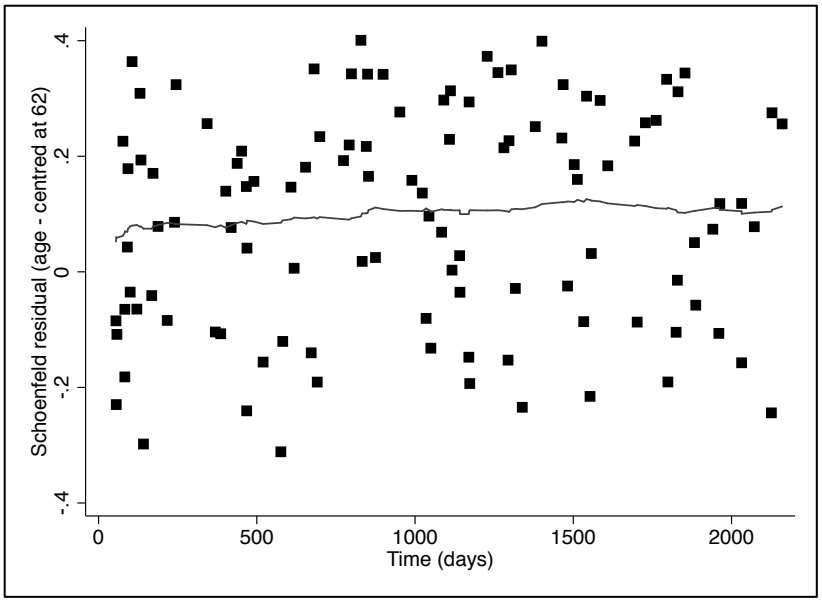

Figure 4. Schoenfeld residual plot of age (centred on 62 years) over the length of study follow-up. A slope of zero indicates proportionality of hazard over time.

CHAPTER V: DISCUSSION

5.1 Population Characteristics

The PLCO study population was selected to attempt to be representative of the general US population in terms of gender distribution and minority representation, while the age range of 55-74 years at recruitment was designed to encompass those at the higher risk of developing one of the PLCO cancers (Prorok et al., 2000).

The selected age range was consistent with what has since been established as "high risk" for lung cancer (Aberle et al., 2011). Furthermore, the average age of those with incident lung cancer cases in the never-smoker sample was 65.6 years (*Table 3*), which is consistent with the reported never-smoker mean of 63.5 years (Nordquist et al., 2004). The age range of this study population is consistent with the lower bounds of the recommended screening age range, while participants are able to age into the upper range (80 years) during follow-up (Humphrey et al., 2013). The PLCO never-smokers are also predominately female (59.7%), however this is consistent with what is known about never-smokers, especially among older populations (Samet et al., 2009).

While minority representation was sought in "appropriate numbers" (Prorok et al., 2000), the never-smoker cohort appears to be disproportionately white. In total, approximately 88% of the study sample identified as a non-Hispanic white, while Asian and black participants represented only 4% of the population each, and Hispanics represented less than 2% (*Table 3*). This differs greatly from the US population, which is approximately 64% white, 16% Hispanic, 4.6% Asian, and 12.3% black (United States Census Bureau, 2013). The under-representation of minorities in the PLCO population

resulted in an inability to properly evaluate racial differences and risk of developing lung cancer, which will be discussed in the forthcoming limitations section.

It has been suggested in previous studies that the PLCO population is more affluent than the general US population (Tammemagi et al., 2011), and this is further demonstrated through the education attainment of this study sample. The PLCO never-smoker population were more likely to have obtained a Masters or Doctoral degree (21% vs. 11% in the US between the ages of 55-74) (United States Census Bureau, 2014), and less likely to have not completed high school (5.7% vs. 18% in the US between the ages of 55-74) (United States Census Bureau, 2014). The misrepresentation of the general population presents potential limitations to the current study, which will be discussed.

5.2 Predictors of Lung Cancer Risk

5.2.1 Age

As was expected, age was a statistically significant predictor of lung cancer risk in never-smokers ($p<0.001$). It has been described as the most consistent and important predictor of LCINS risk, as it represents both an accumulation of lifetime exposures, and an increased likelihood of genetic mutation (McCarthy et al., 2012). The effect per year increase was consistent with what was found in previous lung cancer risk prediction models (HR=1.095 per year, 95% CI 1.057-1.135), indicating a consistent effect in both ever- and never-smokers (Tammemagi et al., 2013). Unlike other models (Tammemagi et al., 2011), the age-risk relationship was most accurately modeled linearly, with the baseline risk centred at the mean age of 62 years.

5.2.2 Body Mass Index

Increasing BMI has previously been shown to have a protective effect on lung cancer risk, however this relationship was suspected to be due to active smokers having a lower BMI (Renehan et al., 2008). In the large meta-analysis by Renehan *et al.*, a relative risk of 0.76 was seen for a 5kg/m^2 increase in BMI, across all smoking strata. However, when limited to never-smokers, this protective effect became lessened, and non-significant. Separate studies have shown a non-significant increase in risk associated with higher BMI in never-smokers, indicating that smoking status may have a confounding effect (Kabat et al., 2007; Renehan et al., 2008).

This study demonstrated a non-significant protective effect of increasing BMI, with a hazard ratio of 0.972 (p=0.184) per one unit increase. While this non-significant effect is consistent with what has been shown in never-smoker populations, it is also nearly identical to the odds ratio produced in the $PLCO_{m2012}$ model (OR=0.973, 95% CI 0.955-0.991), indicating similar BMI-risk relationships between both PLCO ever- and never-smokers (Tammemagi et al., 2013). As with age, nonlinear effects were evaluated, but the most accurate relationship remained linear – consistent with previous models (Tammemagi et al., 2011, 2013).

5.2.3 Education Level

In this study, consistent with other predictive models using the PLCO population, level of educational attainment was used as a surrogate for SES (Tammemagi et al., 2011). Higher SES, and by extension higher educational attainment, has consistently demonstrated protective effects for lung cancer risk compared to individuals with lower

SES/education (Sidorchuk et al., 2009). The effect of increasing education levels was consistent between this study and previous PLCO-derived models, indicating that the effect of achieving a higher education level is independent of smoking status (Tammemagi et al., 2013). A benefit of the PLCO population having a large number of highly educated participants is a large sample size across each of the education groups. This allowed for a more accurate estimation of risk with increasing education level, and thus a better understanding of progressively lower risk with increasing education attainment. However, a drawback of this is that there are fewer than expected low education individuals, providing less accurate estimates of the highest risk groups.

5.2.4 Personal History of Cancer

Given the exclusion criteria of the PLCO study – individuals with prostate, lung, colorectal, or ovarian cancer, as well as those currently undergoing any cancer treatment – there were limitations on participants who had a previous personal history of cancer (National Cancer Institute, 2014; Prorok et al., 2000). As such, the majority of previous cancers in this study were breast cancer. With regards to breast cancer, it is thought that radiation therapy as treatment rather than genetic disposition is the likely cause of future lung cancer (Bevers, 2014).

Findings from this study were consistent with those of Hodgson *et al.* (2014) and Neugut *et al.* (1994), which estimated a relative risk of future lung cancer of between 1.5-3 for women with breast cancer compared to those without. Furthermore, the hazard ratio obtained by this model was consistent with the effect in the $PLCO_{m2012}$ (HR=1.909 compared to OR=1.582 in $PLCO_{m2012}$) (Tammemagi et al., 2013). Previous history of cancer demonstrated the second highest individual risk contribution (behind only multiple

first degree relatives with lung cancer), however this estimate had a wide confidence interval due to a relatively low number of positive cases, which in part is due to the aforementioned PLCO exclusion criteria.

5.2.5 Family History of Lung Cancer

The largest single effect estimate in this model is having two or more first-degree relatives with lung cancer (HR=3.522, 95% CI 0.658-14.354). Family history has been hypothesized to be synergistic with smoking status, as studies have shown an increase effect estimate in ever-smoker and combined populations compared to never-smokers (Lissowska et al., 2010). Family history has been well established as a predictor, both in terms of immediate family members, and through heritable genetic mutations (Brennan et al., 2011). While there are several loci identified that are associated with increased lung cancer risk, these have not been shown to improve prediction beyond family history (Li et al., 2012). Due to the limited availability and practicality of GWAS data, current models limit heritable cancer risk to family history within first-degree relatives, with risk estimates in ever-smoker models consistent with literature (Lissowska et al., 2010; Tammemagi et al., 2013).

The use of a multilevel categorical predictor for family history (no relatives, one relative, two or more relatives) was similar to that by Spitz *et al.* (2007). While the sample size is small at the two or more level, the hazard ratio is significantly higher than the single-relative level, providing evidence of a dose-response relationship regarding LCINS risk.

5.2.6 Chest X-ray in the Past Three Years

Previous CXR serves as a risk factor for lung cancer likely as a result of the diagnostic workup associated with various inflammatory lung diseases, which have been shown to increase the risk of developing lung cancer (Fitzpatrick, 2001). There is little in literature to suggest that receiving an CXR would be a major contributor to lung cancer, with an estimated 0.1% and 0.5% attributable risk of lung cancer associated with a single diagnostic X-ray in males and females respectively (Berrington de González & Darby, 2004). Nonetheless, multilevel categorical predictors regarding CXR exposure have been used in previous risk prediction models (Tammemagi et al., 2011), so there was precedent for inclusion in this study. Unlike previous models, receiving one CXR was combined with no exposure, as there was no risk difference between the two groups. Thus, the only increased risk in never-smokers was associated with receiving multiple CXRs in the three years prior to baseline. This served as a strong predictor, and one of the few significant even at a 5% significance level in the full model.

5.2.7 Living with a Smoker as an Adult

Secondhand smoke exposure has long been established as one of the most consistent predictors of LCINS risk (USDHHS, 2006). The estimated effect of living with a smoker as an adult (household SHS) in this study (HR=1.26, 95% CI 0.56-2.87) fell within the range provided by the United States Surgeon General – a relative risk of between 1.20 and 1.29 (USDHHS, 2006). This means that, despite the high amount of imputed data for this SHS variable, the overall effect estimate remained within a plausible range. One limitation of the PLCO supplemental data collection was the lack of recorded exposure duration for SHS. This meant that the demonstrated three-times

increased risk with heightened exposure – as established in the US Surgeon General's report - was undeterminable in this study, leading to a potential underestimation of SHS household risk for some individuals (USDHHS, 2006).

5.2.8 Suspected Risk Factors Not Included

There were a number of established LCINS risk factors, that have been discussed in depth in literature, that were not included in the final model despite availability in the PLCO dataset. First, SHS exposure in the workplace – which was variable in risk but generally similar to household exposure – demonstrated a protective effect when entered into the full Cox model. This violated one of the conditions for variable selection, the need for a plausible effect consistent with what is already known, and thus was not included. A similar phenomenon was noticed with the lung comorbidity variable, which primarily consisted of cases of COPD. While COPD is thought to occur most often only in ever-smokers, there were cases amongst the never-smokers in the PLCO dataset, and inflammatory lung diseases had been shown to be associated with increased lung cancer risk (Fitzpatrick, 2001). A history of COPD has shown good predictive ability in models of ever-smokers (Tammemagi et al., 2011, 2013), so there was reason to believe it would be associated with increased risk in never-smokers as well. It instead showed a strong protective effect, leading to its removal from the final model for violating one of the necessary inclusion criteria.

A number of other potential predictors were not included due to a failure to improve prediction, a null risk estimate, or a greatly non-significant p-value. Among the variables that fit this description were: gender, race, ibuprofen and aspirin usage, alcohol consumption, and variables pertaining to physical activity levels and dietary intake.

Furthermore, a number of predictors could not be included due to unavailable data in the PLCO dataset. This will be discussed in the limitations section.

5.3 Model Performance

This study developed a model with fair or adequate discrimination (optimism corrected c-statistic= 0.6645, AUC=0.6858), falling short of the cut-point for what is considered good discrimination (Hosmer & Lemeshow, 2000). When compared to the best ever-smoker models developed in the same dataset, $PLCO_{m2011}$ and $PLCO_{m2012}$, this model falls well short (AUC= 0.809 and 0.803 respectively) (Tammemagi et al., 2011, 2013). Despite this, the current model achieves a modest improvement over the adapted $PLCO_{all2014}$ model limited to never-smokers (AUC=0.662) (*Figure 2*) demonstrating that developing a model amongst never-smokers performs better than a successful ever-smoker model including a population of never-smokers (Tammemagi et al., 2014).

The current model performs better than the Spitz model when limited to the never-smoker population (c-statistic= 0.57) (Spitz et al., 2007). Model calibration is worse than in the $PLCO_{all2014}$ model, as well as the original $PLCO_{m2011}$ model (Tammemagi et al., 2011, 2014), but still represents good calibration based on the mean and 90^{th} percentile absolute errors, and non-significant Spiegelhalter's statistic. Despite this, model calibration worsens above a six-year probability of 0.01, indicating that calibration may not be good at any decision-making threshold. This indicates that while it does not achieve the performance of other PLCO models, it does compare well to other never-smoker models and even some ever- and current-smoker models.

5.3.1 Is This Model Suitable for Never-smoker Risk Prediction?

The goal of this thesis was to develop a risk prediction model for never-smokers capable of accurate individual-level risk prediction, similar to the Tammemagi PLCO models (Tammemagi et al., 2014). Which begs the question: does this model achieve the goal of being able to predict lung cancer accurately in never-smokers? According to Harrell (2001), who states an AUC of greater than 0.8 is required for individual-level prediction, this model is not capable.

In a more practical sense, it is important to identify how many never-smokers would be identified for screening using this model, given that none are screened using current criteria (USDHHS, 2006). When the $PLCO_{m2012}$>0.0151 risk probability - which was proven to be more efficient in identifying individuals for screening- is applied to this model, a total of 35 individuals are selected (Tammemagi et al., 2014). This means that out of the 69,272 never-smokers in the PLCO population, only 35 (0.0005%) would be selecting for screening, indicating a very low number of qualifying never-smokers. Furthermore, of these 35 individuals that would be selected for screening based on their six-year predicted risk, none actually developed lung cancer within the first six years of follow-up. Unlike the $PLCO_{all2014}$ model, the highest achieved risk for a never-smoker in this model was 0.0342, however, since these "high-risk" individuals did not get cancer, screening them would in fact be an inefficient use of resources.

Lowering the risk threshold to capture more LCINS cases is not a feasible option, as the vast majority of cases are clustered in the 0-0.4% six-year risk range, along with the majority of the never-smoker population. This clustering of risk scores is likely due to relatively few never-smokers having the traits that most greatly increase risk (multiple

relative family history, multiple CXR exposure, personal history of cancer), and is also likely representative of the true nature of lung cancer risk as exposure to known risk factors is low for never-smokers. Low exposure frequency, coupled with low six-year lung cancer incidence, results in few high-risk individuals and few lung cancers.

5.3.2 Improving LCINS Risk Prediction

The inability of this model, and all never-smoker risk prediction models, to achieve high discrimination leads to the question: what needs to be done to develop a good risk prediction model, and why does the never-smoker population differ from smokers with regards to predictive success? The answer is two-fold; the predictors that were strongest in never-smokers were also fairly rare among the population, leading to few people achieving high risk, and also the lack of a predictor that is as strongly tied to lung cancer risk as smoking is in current and former smokers. Even the strongest predictor in this model (multiple relative family history, HR= 3.522) is much weaker than being a current active smoker, which has been shown to increase lung cancer risk by 10-20-times (Clément-Duchêne et al., 2010).

In order to improve LCINS prediction, it is necessary to identify predictors similar to active smoking. Among the candidates would be a trio of loci identified via GWAS: *5p15, 6p21*, and *15q25* (Brennan et al., 2011; Rudin et al., 2009). The inheritance of risk alleles on these loci has been associated with an 80% increased risk, which is much lower than the risk associated with active smoking (Brennan et al., 2011). To date, including GWAS data in lung cancer models has not improved prediction (Li et al., 2012). While it is very unlikely that any single predictor would improve never-smoker prediction, a combination of GWAS and biomarker data could lead to improved prediction. A

drawback of using GWAS data, as is also the case with clinical biomarker data, is that it requires additional testing or sequencing, and thus reduces the clinical relevancy and usefulness (Tammemagi, 2015). An appeal to this model, along with others (Tammemagi et al., 2013), is that they can be carried out by public health officials using readily available data, or data that can easily be obtained from individuals.

5.4 Limitations

5.4.1 Unavailable Variables

While there were many suspected risk factors for LCINS described in the literature, many of the strongest ones were not available in the PLCO dataset. Missing information that could potentially be valuable includes environmental exposures, occupational exposures, and residential radon exposure. There is some precedent for the use of these exposures in lung cancer predictive modeling, as Spitz *et al.* used self-reported history for select exposures while both Spitz *et al.* and Cassidy *et al.* attempted to ascertain asbestos exposure through documented workplace history (Cassidy et al., 2008; Spitz et al., 2007). However – as the authors report – these variables were self-reported and not validated, and thus subject to misclassification bias (Spitz et al., 2007). The asbestos exposure variable used in the model by Cassidy *et al.* was complex and could not be utilized in practice. Furthermore, developing a reliable "occupational risk" variable could be a very time consuming process, with a nearly unlimited number of potential jobs that would have to be evaluated for potential exposure risk, and categorized accordingly. While it would be difficult to develop, an accurate and easily categorized occupational exposure variable may have been beneficial to model performance. Furthermore, while environmental and residential radon exposures are difficult to

measure at an individual level, ecologic measures (urban vs. rural, basement dwelling vs. upper-storey, etc.) could have allowed for some investigation of these risk factors. A drawback of using these ecologic measures is that these variables would be prone to misclassification, and likely provide little utility or in risk prediction or potentially worsen predictive performance. Finally, given the restrictions on PLCO inclusion criteria (no previous history of PLCO cancers), the risk associated with some of the most common types of cancer could not be determined (Prorok et al., 2000). With 13.2% of second primary cancers occurring in the same tissue as the original cancer, it is possible that having previous lung cancer as a predictor would have strengthened model performance (Bevers, 2014).

Expanded data on some variables already included in the dataset, such as more detailed dose-related SHS exposures, would also have allowed a better understanding of individual-level risk. As was mentioned in the US Surgeon General's report, there was a noted dose-response relationship (3x increased risk associated with >4 hours per day exposure compared to no exposure) (USDHHS, 2006). Being able to better elucidate this relationship among the population could have added to the predictive performance of the model by providing a higher-risk group, perhaps further separating cases from non-cases.

5.4.2 Population Representation

As was mentioned in *Section 5.1*, the PLCO study sample was not completely representative of the United States population as a whole. The study sample is predominately white and more educated than would be expected by truly random sampling (United States Census Bureau, 2013, 2014). As is outlined by Murthy *et al.* (2004), underrepresentation of minorities and lower SES individuals is a recurrent

problem in cancer screening trials, likely a product of both selective recruitment and a mistrust in the health care system (Murthy et al., 2004). Minimal representation of minorities eliminated the possibility of doing race-specific subset models. Given the modest performance of this model, the supposed heightened risk of some racial groups, and the evidence of good subpopulation-specific versions of other models, it would have been beneficial to investigate in this model (Haiman et al., 2006; Tammemagi et al., 2011; Thun et al., 2008).

5.4.3 Self-reported Data

PLCO data were obtained via self-reported questionnaire, which leads to questions about the validity of the responses (Prorok et al., 2000). The responses to this questionnaire, especially those pertaining to smoking status, alcohol consumption, dietary choices, and physical activity, are subject to social desirability bias (van de Mortel, 2008). Since participants may be more inclined to respond in a way reflecting societal norms or ideals, smoking status and alcohol consumption may be underestimated while physical activity and dietary consumption may appear more favourable (van de Mortel, 2008). This could have led to a washing out of the effects of predictors that were subject to social desirability bias, and worsening of model prediction as the true effect of these variables on lung cancer risk would not be modeled.

5.5 Strengths

This study had a number of strengths, first of which is the use of the PLCO dataset. This is a large, prospective study which contains a large number of never-smokers. It avoids the limitations of some other lung cancer models, which use case-

control data and are often matched based on age, gender, and smoking status (Cassidy et al., 2008; Spitz et al., 2007). Using prospective data allowed for use of incidence data, which is ideal for risk prediction modeling (Adami et al., 2008; Tammemagi et al., 2011). These data are also high-quality, with rigorous follow-up procedures ensuring that data were available for nearly all individuals throughout the study process (Prorok et al., 2000; Tammemagi et al., 2011). This is the same data source used in other high-performing lung cancer risk prediction models, further indicating the usefulness of these data (Tammemagi et al., 2011, 2013).

Secondly, advanced statistical techniques were used throughout the methodological process. First, using multiple imputation allowed for an accurate means of estimating missing data, avoided losing observations and outcomes due to the use of Supplemental and Dietary data, and thus allowed for the inclusion of SHS and potential inclusion of other predictors in the final model. Next, using Cox proportional hazards regression allowed for the inclusion of time-to-event data, maximizing the amount of information that went into generating risk scores and allowing for multiple follow-up periods to be used – although only six-year risk was ultimately included in the final model. The use of MFPs allowed for the investigation of non-linear relationships of continuous variables, and while the relationship remained linear in the final model, it was not simply assumed to be so. Bootstrapping internal validation techniques allowed for optimism correction and providing a more realistic estimate of the c-statistic, while also allowing for this study to incorporate the full sample into model-building without reserving separate training and validation samples. By not adhering to a strict p-value cut-point (<0.05), and utilizing intelligent variable selection for model building, we avoided

missing potentially useful predictors while also ensuring predictors with spurious relationships (such as the COPD-risk relationship) were not included in the final model. The final model is parsimonious, containing only seven predictors, and thus should be easily reproducible if validated in other studies (Harrell et al., 1984; Moons et al., 2012)

This study attempted to expand and improve on previous models by investigating a number of novel risk factors, based on what is established or suspected in LCINS literature. While no predictors were unique specifically to this model, relationships between ibuprofen and aspirin use, alcohol consumption, physical activity, female hormone usage, and dietary factors – among others – were investigated for their potential role in predicting lung cancer. Many of these variables were made possible by the inclusion of Supplemental and Dietary Questionnaire data, and the use of multiple imputations allowed these variables to be used without a reduction in sample size. Most never-smoker models were developed alongside ever- and current-smoker models (Spitz et al., 2007) or adapted from existing smoker-developed models (Tammemagi et al., 2014), and thus this model benefits by being developed in a large population of exclusively never-smokers, and using predictors determined from an LCINS-specific literature review.

5.6 Implications

Based on the adequate discrimination, and the fact that no lung cancer cases were detected amongst the few high-risk individuals identified, this suggests that the current model is not suitable for clinical use or individual-level risk prediction in never-smokers. Without an overwhelmingly strong predictor, such as smoking intensity and duration in smokers, never-smoker prediction appears to be limited to moderate discrimination and

population-level prediction at best. This is not to say that never-smokers should not be eligible for screening, as 10-15% represents a large amount of lung cancer cases overall, only that there needs to be better, more efficient methods for identifying those never-smokers at high risk.

5.7 Future Research

The most important next step in LCINS research is to identify predictors, similar to smoking status, that allow for improved discrimination. A likely candidate is the use of genomic data, particularly a select group of genes associated with increased risk amongst never-smokers. In addition, biomarkers such as haemoglobin and fasting glucose levels have been included in lung cancer prediction models (Tammemagi, 2015), and therefore it may be beneficial to investigate their performance in LCINS prediction. However, it is also important to be mindful of the practical utility when developing future LCINS prediction models, as genomic and biomarker data are not as readily available. Further, more practical, areas of future research include the investigation of subpopulations within the never-smoker population, particularly specific race groups that may be at higher risk. Asian populations have demonstrated a much higher incidence of LCINS than North America, or individuals of Asian decent living in North America (Thun et al., 2008), so developing a model in this population is of public health importance and it may be possible to develop a good prediction model in Asian never-smokers. This could allow for the identification of at least some high-risk groups, and potentially lead to screening, early detection, and reduced mortality for some of the 10 to 15% of lung cancer cases that occur in never-smokers.

5.8 Conclusion

With the success of LDCT screening in improving detection and diagnosis of lung cancer (Aberle et al., 2011), specifically in high-risk individuals, it has become paramount to develop ways of detecting exactly who constitutes "high-risk". To date, lung cancer risk prediction models have demonstrated a good ability to identify high-risk individuals, with recent studies showing more efficient screening identification than the United States guidelines (Tammemagi et al., 2014). However, a common theme amongst prediction models and screening guidelines is the exclusion of never-smokers from high-risk groups, despite an estimated 10-15% of lung cancer cases occurring in never-smokers (Samet et al., 2009). Never-smokers are subject to a different set of lung cancer risk factors than their ever-smoking counterparts, and thus require a specifically developed model to allow for accurate risk prediction (Sun et al., 2007; Tammemagi et al., 2014). Therefore, the aim of this thesis was to develop and validate an accurate risk prediction model in never-smokers, with the goal of achieving individual level predictive performance.

This current model was developed using the same population, and many of the same statistical techniques as other high-performing lung cancer models (Tammemagi et al., 2011). However, this model does not achieve the same level of predictive performance, achieving good calibration and fair to adequate discrimination, and is therefore not appropriate for individual-level risk prediction. Despite rigorous statistical methodology and the use of high-quality data, this study was not able to develop a model in never-smokers with the same capability of ever- and current-smoker models. This

study does represent an improvement on existing never-smoker models (Spitz et al., 2007; Tammemagi et al., 2014) indicating progress in the field of LCINS risk prediction.

REFERENCES

Aberle, D., Adams, A., Berg, C., Black, W., Clapp, J., Fagerstrom, R., … Sicks, J. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, *365*(5), 395–409.

Adami, H., Hunter, D., & Trichopoulos, D. (2008). *Textbook of cancer epidemiology* (2nd Editio, p. 748). New York, New York: Oxford University Press.

Alberg, A. J., Brock, M. V, Ford, J. G., Samet, J. M., & Spivack, S. D. (2013). Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, *143*(5 Suppl), e1S–29S. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23649439

American Cancer Society. (2006). Cancer Facts & Figures, 2006: Special Edition: Environmental Pollutants and Cancer. Retrieved from www.cancer.org

American Cancer Society. (2014). *Lung Cancer (Non-Small Cell) What is cancer ?*. *American Cancer Society*.

American Joint Committee on Cancer. (2010). *Cancer staging manual*. (S. B. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, & A. Trotti III, Eds.) (7th ed., p. 648). Chicago, IL: Springer.

Bach, P. B., Kattan, M. W., Thornquist, M. D., Kris, M. G., Tate, R. C., Barnett, M. J., … Begg, C. B. (2003). Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute*, *95*(6), 470–8.

Baik, C. S., Strauss, G. M., Speizer, F. E., & Feskanich, D. (2010). Reproductive factors, hormone use and risk of lung cancer in postmenopausal women, the Nurses' Health Study. *Cancer Epidemiology Biomarkers Prevention*, *19*(10), 2525–33.

Bell, N., Dickinson, J., & Singh, H. (2014). *Protocol : Screening for Lung Cancer* (pp. 1–10).

Berman, D. W., & Crump, K. S. (2008). A meta-analysis of asbestos-related cancer risk that addresses fiber size and mineral type. *Critical Reviews in Toxicology*, *38 Suppl 1*, 49–73. http://doi.org/10.1080/10408440802273156

Berrington de González, A., & Darby, S. (2004). Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *Lancet*, *363*(9406), 345–51. http://doi.org/10.1016/S0140-6736(04)15433-0

Besaratinia, A., & Pfeifer, G. P. (2008). Second-hand smoke and human lung cancer. *The Lancet. Oncology*, *9*(7), 657–66.

Bevers, T. B. (2014). Screening for second primary cancers. In L. E. Foxhall & M. A. Rodriguez (Eds.), *Advances in Cancer Survivorship Management* (pp. 299–321). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-0986-5

Black, W. C., Gareen, I. F., Soneji, S. S., Sicks, J. D., Keeler, E. B., Aberle, D. R., … Gatsonis, C. (2014). Cost-Effectiveness of CT Screening in the National Lung Screening Trial. *New England Journal of Medicine*, *371*(19), 1793–1802. http://doi.org/10.1056/NEJMoa1312547

Blattenberger, G., & Lad, F. (1985). Separating the Brier score into calibration and refinement separating a graphical exposition components: a graphical exposition. *The American Statistician*, *39*(1), 26–32.

Brennan, P., Hainaut, P., & Boffetta, P. (2011). Genetics of lung-cancer susceptibility. *The Lancet. Oncology*, *12*(4), 399–408. http://doi.org/10.1016/S1470-2045(10)70126-1

Brenner, D. R., McLaughlin, J. R., & Hung, R. J. (2011). Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *PloS One*, *6*(3), e17479.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 2–4.

Canadian Cancer Society. (2014). Signs and symptoms of lung cancer. Retrieved August 20, 2014, from http://www.cancer.ca/en/cancer-information/cancer-type/lung/signs-and-symptoms/?region=bc

Carney, D. (1995). Lung Cancer Biology. *Seminars in Radiation Oncology*, *5*(1), 4–10.

Cassidy, A., Myles, J. P., van Tongeren, M., Page, R. D., Liloglou, T., Duffy, S. W., & Field, J. K. (2008). The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer*, *98*(2), 270–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2361453&tool=pmcentrez&rendertype=abstract

Centers for Medicare and Medicaid Services. (2015). Decision memo for screening for lung cancer with low dose computed tomography.

Clément-Duchêne, C., Vignaud, J.-M., Stoufflet, A., Bertrand, O., Gislard, A., Thiberville, L., … Paris, C. (2010). Characteristics of never smoker lung cancer including environmental and occupational risk factors. *Lung Cancer*, *67*(2), 144–50.

Cleves, M. A., Gould, W. W., Gutierrez, R. G., & Marchenko, Y. U. (2008). *An introduction to survival analysis using Stata* (Second Edi, p. 372). College Station, Texas: Stata Press.

Couraud, S., Zalcman, G., Milleron, B., Morin, F., & Souquet, P.-J. (2012). Lung cancer in never smokers--a review. *European Journal of Cancer*, *48*(9), 1299–311.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Damjanov, I. (2012). *Pathology for the Health Professions* (Fourth Ed, p. 526). St. Louis, Missouri: Elsevier Inc.

De Stefani, E., Boffetta, P., Deneo-Pellegrini, H., Mendilaharsu, M., Carzoglio, J. C., Ronco, A., & Olivera, L. (1999). Dietary antioxidants and lung cancer risk: a case-control study in Uruguay. *Nutrition and Cancer*, *34*(1), 100–110.

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011). Lung Cancer: epidemiology, etiology and prevention. *Clinical Chest Medicine*, *32*(4), 1–61. http://doi.org/10.1016/j.ccm.2011.09.001.Lung

Endo, H., Yano, M., Okumura, Y., & Kido, H. (2014). Ibuprofen enhances the anticancer activity of cisplatin in lung cancer cells by inhibiting the heat shock protein 70. *Cell Death & Disease*, *5*(1), e1027. http://doi.org/10.1038/cddis.2013.550

Fitzpatrick, F. A. (2001). Inflammation, carcinogenesis and cancer. *International Immunopharmacology*, *1*, 1651–1667.

Goodman, G. E., Thornquist, M. D., Balmes, J., Cullen, M. R., Meyskens, F. L., Omenn, G. S., … Williams, J. H. (2004). The Beta-Carotene and Retinol Efficacy Trial: incidence of lung cancer and cardiovascular disease mortality during 6-year follow-up after stopping beta-carotene and retinol supplements. *Journal of the National Cancer Institute*, *96*(23), 1743–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15572756

Haiman, C. A., Stram, D. O., Wilkens, L. R., Pike, M. C., Kolonel, L. N., Henderson, B. E., & Le Marchand, L. (2006). Ethnic and racial differences in the smoking-related risk of lung cancer. *New England Journal of Medicine*, *354*(4), 333–342.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating (ROC) characteristic curve. *Radiology*, *143*(1), 29–36.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, New York: Springer.

Harrell, F. E. (2014). Regression Modelling Strategies.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. a. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, *3*(2), 143–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6463451

Hocking, W. G., Hu, P., Oken, M. M., Winslow, S. D., Kvale, P. a, Prorok, P. C., … Berg, C. D. (2010). Lung cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. *Journal of the National Cancer Institute*, *102*(10), 722–31. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873186&tool=pmcentrez&rendertype=abstract

Hodgson, D. C., Ng, A., & Travis, L. B. (2014). Radiation-related second primary cancers: clinical perspectives. In P. Rubin, L. S. Constine, & L. B. Marks (Eds.), *ALERT - Adverse Late Effects of Cancer Treatment* (pp. 241–252). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-72314-1

Hoffman, P. C., Mauer, a M., & Vokes, E. E. (2000). Lung cancer. *Lancet*, *355*, 479–85.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second edi, p. 375). John Wiley & Sons, Inc.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., … EJ, C. K. (2011). *SEER Cancer statistics review 1975-2011*. Bethesda, MD. Retrieved from http://seer.cancer.gov/csr/1975_2011/

Hu, J., Mao, Y., Dryer, D., & White, K. (2002). Risk factors for lung cancer among Canadian women who have never smoked. *Cancer Detection and Prevention*, *26*(2), 129–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12102147

Humphrey, L. L., Deffeback, M., Pappas, M., Baumann, C., Artis, K., Mitchell, J. P., … Statore, C. G. (2013). Screening for lung cancer with low-dose computed tomography: a systematic review to update the U.S preventative services task force recommendation. *Annals of Internal Medicine*, *159*(6).

International Agency for Research on Cancer. (2012). *Indoor emissions from household combustion of coal* (Vol. 2006).

Kabat, G. C., Miller, A. B., & Rohan, T. E. (2007). Body mass index and lung cancer risk in women. *Epidemiology*, *18*(5), 607–12. Retrieved from http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-200709000-00014

Krewski, D., Lubin, J. H., Zielinski, J. M., Alavanja, M., Catalan, V. S., Field, R. W., … Wilcox, H. B. (2005). Residential radon and risk of lung cancer. *Epidemiology*, *16*(2), 137–45. Retrieved from

http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-200503000-00001

Lagarde, F., Axelsson, G., Damber, L., Mellander, H., Nyberg, F., & Pershagen, G. (2001). Residential radon and lung cancer among never-smokers in Sweden. *Epidemiology*, *12*(4), 396–404.

Li, H., Yang, L., Zhao, X., Wang, J., Qian, J., Chen, H., … Lu, D. (2012). Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. *BMC Medical Genetics*, *13*(1), 118. http://doi.org/10.1186/1471-2350-13-118

Lissowska, J., Foretova, L., Dabek, J., Zaridze, D., Szeszenia-Dabrowska, N., Rudnai, P., … Boffetta, P. (2010). Family history and lung cancer risk: international multicentre case-control study in Eastern and Central Europe and meta-analyses. *Cancer Causes & Control*, *21*(7), 1091–104. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20306329

Matakidou, A., Eisen, T., & Houlston, R. S. (2005). Systematic review of the relationship between family history and lung cancer risk. *British Journal of Cancer*, *93*(7), 825–33. http://doi.org/10.1038/sj.bjc.6602769

McCarthy, W. J., Meza, R., Jeon, J., & Moolgavkar, S. H. (2012). Chapter 6: Lung cancer in never smokers: epidemiology and risk prediction models. *Risk Analysis*, *32 Suppl 1*, S69–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3485693&tool=pmcentrez&rendertype=abstract

Meza, R., Hazelton, W. D., Colditz, G. A., & Moolgavkar, S. H. (2008). Analysis of lung cancer incidence in the nurses ' health and the health professionals ' follow-up studies using a multistage carcinogenesis model. *Cancer Causes & Control*, *19*(3), 317–328.

Midthun, D. E. (2013). Early diagnosis of lung cancer. *F1000prime Reports*, *5*, 12. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3616602&tool=pmcentrez&rendertype=abstract

Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart (British Cardiac Society)*, *98*(9), 683–90. http://doi.org/10.1136/heartjnl-2011-301246

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: Part I. Two-state situation. *Journal of Applied Meteorology*, *11*.

Murthy, V. H., Krumholz, H. M., & Gross, C. P. (2004). Participation in clinical trials: race-, sex- and age-based disparities. *JAMA*, *291*(22), 2070–2076.

Muscat, J. E., & Wynder, E. L. (1995). Lung cancer pathology in smokers, ex-smokers and never smokers. *Cancer Letters*, *88*(1), 1–5.

National Cancer Institute. (2013). Cancer staging. Retrieved August 18, 2014, from http://www.cancer.gov/cancertopics/factsheet/detection/staging

National Cancer Institute. (2014). Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Retrieved March 28, 2014, from http://prevention.cancer.gov/plco

Neugut, A., Ph, D., Murray, T., Santos, J., & Robinson, E. (1994). Increased risk of kung cancer after breast cancer radiation therapy in cigarette smokers. *Cancer*, *73*(6).

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine*, *16*, 611–626.

Nordquist, L. T., Simon, G. R., Cantor, A., Alberts, W. M., & Bepler, G. (2004). Improved survival in never-smokers vs current smokers with primary adenocarcinoma of the lung. *Chest*, *126*(2), 347–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15302716

Nyberg, F., Argenius, V., Svartengren, K., Svensson, C., & Pershagen, G. (1998). Dietary factors and risk of lung cancer in never-smokers. *International Journal of Cancer*, *436*, 430–36.

Oken, M. M., Hocking, W. G., Kvale, P. a, Andriole, G. L., Buys, S. S., Church, T. R., … Berg, C. D. (2011). Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA*, *306*(17), 1865–73. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22031728

Omenn, G. S., Goodman, G. E., Thornquist, M. D., Balmes, J., Cullen, M. R., Glass, a, … Hammar, S. (1996). Risk factors for lung cancer and for intervention effects in CARET, the Beta-Carotene and Retinol Efficacy Trial. *Journal of the National Cancer Institute*, *88*(21), 1550–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8901853

Pinsky, P. F., Miller, a, Kramer, B. S., Church, T., Reding, D., Prorok, P., … Berg, C. D. (2007). Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *American Journal of Epidemiology*, *165*(8), 874–81. http://doi.org/10.1093/aje/kwk075

Prorok, P. C., Andriole, G. L., Bresalier, R. S., Buys, S. S., Chia, D., Crawford, E. D., … Weissfeld, J. L. (2000). Design of the Prostate , Lung , Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Controlled Clinical Trials*, *2456*, 273–309.

Ramanakumar, A. V, Parent, M.-E., & Siemiatycki, J. (2007). Risk of lung cancer from residential heating and cooking fuels in Montreal, Canada. *American Journal of Epidemiology*, *165*(6), 634–42. http://doi.org/10.1093/aje/kwk117

Reid, B. C., Ghazarian, A. A., Demarini, D. M., Sapkota, A., Jack, D., & Lan, Q. (2012). Research opportunities for cancer associated with indoor air pollution from solid-fuel combustion. *Environmental Health Perspectives*, *120*(11), 1495–98.

Renehan, A. G., Tyson, M., Egger, M., Heller, R. F., & Zwahlen, M. (2008). Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet*, *371*(9612), 569–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18280327

Royston, P., Moons, K. G. M., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research : Developing a prognostic model. *BMJ*, *338*, 1373–1377. http://doi.org/10.1136/bmj.b604

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (p. 258). New York, New York: John Wiley & Sons, Inc.

Rudin, C. M., Avila-Tang, E., Harris, C. C., Herman, J. G., Hirsch, F. R., Pao, W., … Samet, J. M. (2009). Lung cancer in never smokers: molecular profiles and therapeutic implications. *Clinical Cancer Research*, *15*(18), 5646–61. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2950319&tool=pmcentrez&rendertype=abstract

Sagerup, C. M. T., Småstuen, M., Johannesen, T. B., Helland, Å., & Brustugun, O. T. (2010). Sex-specific trends in lung cancer incidence and survival: a population study of 40,118 cases. *Thorax*, *66*(4), 301–7. http://doi.org/10.1136/thx.2010.151621

Samet, J. M., Avila-Tang, E., Boffetta, P., Hannan, L. M., Olivo-Marston, S., Thun, M. J., & Rudin, C. M. (2009). Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clinical Cancer Research*, *15*(18), 5626–5645.

Santillan, A. A., Camargo, C. A., & Colditz, G. A. (2003). A meta-analysis of asthma and risk of lung cancer (United States). *Cancer Causes & Control*, *14*, 327–34.

Sharma, S. V, Bell, D. W., Settleman, J., & Haber, D. a. (2007). Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews. Cancer*, *7*(3), 169–81. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17318210

Shigematsu, H., Lin, L., Takahashi, T., Nomura, M., Suzuki, M., Wistuba, I. I., … Gazdar, A. F. (2005). Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *Journal of the National Cancer Institute*, *97*(5), 339–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15741570

Sidorchuk, A., Agardh, E. E., Aremu, O., Hallqvist, J., Allebeck, P., & Moradi, T. (2009). Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis. *Cancer Causes & Control*, *20*(4), 459–71. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19184626

Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., … Etzel, C. J. (2007). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17470739

Statistics Canada. (2015). *Canadian Cancer Statistics Special topic : Predictions of the future burden of cancer in Canada*.

Steyerberg, E. W. (2009). *Clinical Prediciton Models: A practical approach to development, validation and updating*. (M. Gail, A. Tsiatis, K. Krickeberg, W. Wong, & J. Sarnet, Eds.) (p. 497). New York, New York: Springer.

Subramanian, J., & Govindan, R. (2007). Lung cancer in never smokers: a review. *Journal of Clinical Oncology*, *25*(5), 561–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17290066

Subramanian, J., Velcheti, V., Gao, F., & Govindan, R. (2007). Presentation and stage-specific outcomes of lifelong never-smokers with non-small cell lung cancer (NSCLC). *Journal of Thoracic Oncology*, *2*(9), 827–830.

Sun, S., Schiller, J. H., & Gazdar, A. F. (2007). Lung cancer in never smokers--a different disease. *Nature Reviews: Cancer*, *7*(10), 778–90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17882278

Tammemagi, M. C. (2015). Application of risk prediction models to lung cancer screenning: a review. *Journal of Thoracic Imaging*, *30*(2).

Tammemagi, M. C., Church, T. R., Hocking, W. G., Silvestri, G. a, Kvale, P. a, Riley, T. L., … Berg, C. D. (2014). Evaluation of the Lung Cancer Risks at Which to Screen Ever- and Never-Smokers: Screening Rules Applied to the PLCO and NLST Cohorts. *PLoS Medicine*, *11*(12), e1001764. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25460915

Tammemagi, M. C., Katki, H. a, Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. a, … Berg, C. D. (2013). Selection criteria for lung-cancer screening. *The New*

England Journal of Medicine, *368*(8), 728–36. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3929969&tool=pmcentrez&rendertype=abstract

Tammemagi, M. C., Neslund-Dudas, C., Simoff, M., & Kvale, P. (2004). Smoking and lung cancer survival: The role of comorbidity and treatment. *Chest*, *125*(1), 27–37.

Tammemagi, M. C., Pinsky, P. F., Caporaso, N. E., Kvale, P. a, Hocking, W. G., Church, T. R., … Prorok, P. C. (2011). Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. *Journal of the National Cancer Institute*, *103*(13), 1058–1068. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3131220&tool=pmcentrez&rendertype=abstract

Thun, M. J., Hannan, L. M., Adams-Campbell, L. L., Boffetta, P., Buring, J. E., Feskanich, D., … Samet, J. M. (2008). Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Medicine*, *5*(9), e185. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2531137&tool=pmcentrez&rendertype=abstract

Thun, M. J., Henley, S. J., Burns, D., Jemal, A., Shanks, T. G., & Calle, E. E. (2006). Lung cancer death rates in lifelong nonsmokers. *Journal of the National Cancer Institute*, *98*(10), 691–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16705123

Toh, C., Gao, F., Lim, W.-T., Leong, S.-S., Fong, K.-W., Yap, S.-P., … Tan, E.-H. (2006). Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *24*(15), 2245–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16710022

United States Census Bureau. (2013). *Annual estimates of the resident population by sex, race, and Hispanic origin for the United States, states, and counties: April 1, 2010 to July 1, 2013*.

United States Census Bureau. (2014). *Educational attainment of the population 18 years and over, by age, sex, race, and hispanic origin: 2014*.

United States Department of Health and Human Services. (2006). *The Health Consequences of Involuntary Exposure to Tobacco Smoke A Report of the Surgeon General*. Atlanta, Georgia: US Department of Health and Human Services, Centres for Disease Control and Prevention,Coordinating Centre for Health Promotion,National Centre for Chronic Disease and Prevention and Health Promotion, Office on Smoking and Health.

US National Research Council. (1999). *Biological effects of ionizing radiation (BEIR) VI Report: "The health effects of exposure to indoor radon."*

Vach, W. (2013). *Regression models as a tool in medical research* (First Edit, p. 473). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Van de Mortel, T. F. (2008). Faking it : social desirability response bias in self- report research report research. *Australian Journal of Advanced Nursing*, *25*(4), 40–48.

Van Klaveren, R. J. (2011). Lung cancer screening. *European Journal of Cancer*, *47 Suppl 3*, S147–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21943968

Van Meerbeeck, J. P., Fennell, D. A., & De Ruysscher, D. K. M. (2011). Small-cell lung cancer. *Lancet*, *378*, 1741–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21565397

Villeneuve, P. J., Parent, M.-É., Harris, S. a, & Johnson, K. C. (2012). Occupational exposure to asbestos and lung cancer in men: evidence from a population-based case-control study in eight Canadian provinces. *BMC Cancer*, *12*, 595. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3534484&tool=pmcentrez&rendertype=abstract

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, *165*(6), 710–8. http://doi.org/10.1093/aje/kwk052

Voorrips, L. E., Goldbohm, R. A., Verhoeven, D. T. H., van Poppel, G. A. F. C., Sturmans, F., Hermus, R. J. J., & van den Brandt, P. A. (2000). Vegetable and fruit consumption in the Netherlands Cohort Study on Diet and Cancer. *Cancer Causes & Control*, *11*(2), 101–115.

Wakelee, H. A., Chang, E. T., Gomez, S. L., Keegan, T. H., Feskanich, D., Clarke, C. A., … West, D. W. (2007). Lung cancer incidence in never smokers. *Journal of Clinical Oncology*, *25*(5), 472–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2764546&tool=pmcentrez&rendertype=abstract

Ward, E., Jemal, A., Cokkinides, V., Singh, G. K., Cardinez, C., Ghafoor, A., & Thun, M. (2004). Cancer disparities by race/ethnicity and socioeconomic status. *A Cancer Journal for Clinicians*, *54*(2), 78–93.

White, I. R., Royston, P., & Wood, A. M. (2009). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–99. http://doi.org/10.1002/sim.4067

Willett, W. C., & Trichopoulos, D. (1996). Nutrition and cancer: a summary of the evidence. *Cancer Causes & Control*, *7*(1), 178–180.

Wood, D. E., Eapen, G. A., Ettinger, D. S., Hou, L., Kazerooni, E., Klippenstein, D., … Yang, S. C. (2012). Lung Cancer Screening. *Journal of the National Comprehensive Cancer Network*, *10*, 240–65.

Woodward, M. (2013). *Epidemiology: Study design and data analysis* (3rd Editio, p. 898). Boca Raton, FL: CRC Press.

World Health Organization. (2004). *Pathology & Genetics: Tumours of the Lung, Pleura, Thymus and Heart*. (W. D. Travis, E. Brambilla, H. K. Muller-Hermelink, & C. C. Harris, Eds.) (p. 344). Geneva, Switzerland: World Health Organization.

Yang, P. (2011). Lung cancer in never smokers. *Semin Respir Crit Care Med*, *32*(1), 10–21.

Yao, Y., Gu, X., Zhu, J., Yuan, D., & Song, Y. (2013). Hormone replacement therapy in females can decrease the risk of lung cancer: a meta-analysis. *PloS One*, *8*(8), e71236. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3743826&tool=pmcentrez&rendertype=abstract

Zhou, W., & Christiani, D. C. (2011). East meets west: ethnic differences in epidemiology and clinical behaviors of lung cancer between East Asians and Caucasians. *Chinese Journal of Cancer*, *30*(5), 287–92.